PROCEEDINGS
OF THE
NATIONAL ACADEMY OF SCIENCES
Volume 23 April 15, 1937 Number 4

IONIZATION THEORY AND RADIO-BIOLOGICAL REACTIONS

By Carl E. Nurnberger

Department of Radiology, Peiping Union Medical College

Communicated February 2, 1937

Some recent investigations of the biological effects of neutrons indicate "selective effects" which are more pronounced than any observed for other penetrating radiations. Lawrence and Lawrence[1] have found that per unit of ionization in live rats neutrons are possibly 5 times more effective than x-rays in altering the blood picture of the rats; Lawrence, Aebersold and Lawrence[2] have shown that the dose of x-rays necessary to kill mouse sarcoma 180 *in vitro* is about 3.5 times the lethal dose for the mouse, while if neutrons are used, a dose equal to 2.9 times the lethal dose is necessary to kill the tumor; and Zirkle and Aebersold[3] concluded that one roentgen of neutrons is 10 times more effective than a roentgen of x-rays in retarding the growth of wheat seedlings. Particular attention should be given to the fact that the radiation dose in each of these experiments is expressed in terms of ionization either in air or in the biological materials themselves. The latter was deduced from measurements on methane gas. A simple interpretation of the results is that an ion in a biological system generated by a neutron is more active in exhibiting certain effects than an ion produced by x-rays in the same system. Moreover, the differential effects of sarcoma and normal tissue are greater for neutron ions than for x-ray ions. These observations have considerable significance in the possible use of neutrons for the treatment of malignant diseases. Their interpretation raises important questions about ionization and its relationship to radio-biological reactions in general.

At the present time experience with radio-biological problems is restricted almost entirely to the initial administration of a known physical quantity of radiation and to the final observation of some gross changes in the irradiated materials. The series of reactions which occur between these events are not understood; they are complex in nature and many in number. But, in general, the reactions are thought to be, in the order named, ionization, physical-chemical transformations and disturbances

of physiological-biological functions. In other words, the absorbed radiant energy is dissipated, probably by the decomposition of the H_2O in the tissues, through the formation of positive and negative ions. The transference of energy from the ions to the cell constituents instigates chemical and physical transformations in the cell structures, and finally, as a result of these transformations, the normal activity of the cells is either altered or destroyed. Immediately, numerous questions arise. Are the "selective effects" merely apparent (the ionization in the tissues may be deducted incorrectly), or are they real, and if real, how do they depend upon the primary reaction of ionization and the secondary reactions of chemical transformations? Existing experimental data are too meager to attempt any sort of satisfactory answer. Some portions of the problem, however, are worthy of discussion now.

The phenomenon of ionization is not new; in fact, the application of the ionization theory to radio-chemical reactions in gases has attained much success.[4] On the other hand, the extension of the theory to liquids and solids is limited to only a few investigations, notwithstanding the fact that many of the deductions from studies of gases are logically applicable to other states of matter. The ionization of gases by x-rays and neutrons is a transformation of energy into new types. A major part of the x-ray energy goes into kinetic energy of ejected photo-electrons or recoil electrons; whereas the energy of neutrons is transformed to kinetic energy of ejected protons or recoil nuclei. The moving electrons and protons dissipate their energy in the formation of ions. The ions in a specific gas are identical regardless of their origin. These concepts for the most part are believed to be true for other states of matter also. Neutron ionization in certain liquids and solids may not result entirely from ejected protons. In substances which contain elements highly absorptive of neutrons, for example, some of the rare earths, an appreciable part of the total ionization may result from rays emitted by elements made radioactive. But biological materials do not contain significant quantities of these elements; hence, hydrogen must be the principal absorber of the neutrons and the ejected protons produce the major portion of the ionization. The chief measurable distinction between proton and electron ionization is the distribution of ions in the paths of the rays.* Protons, being much heavier than electrons, rapidly lose their kinetic energy by intense ionization along their paths. Their range is shorter than the range of high velocity electrons. Although the latter suffer many collisions with heavy molecules, less energy is lost at each collision. The density of ions per unit proton path is more than 100 times greater than the number of ions in an equal length of path of electrons generated by x-rays. Assuming that ionic concentration is the principal differentiating factor of proton and electron ionization, and

that chemical and physical changes result from ionization, then it appears worth while to investigate liquid systems which are chemically less complex than the semi-solid media of living tissues and wherein the chemical yields per ion pair can be compared more accurately. Fortunately, neutron irradiation is not necessary in order to test this point, for α-particles also produce ionic concentrations more than 100 times greater than electron ionization. Therefore, a comparison of the chemical effects of α-particles and x-rays in the same system should give some indication of the influence of ionic concentrations.

The calculated values of the chemical yields per ion-pair in experiments with tyrosine are different for x-ray and α-ray irradiation. In 1928 Stenstrom and Lohmann[5] irradiated aqueous solutions of tyrosine with 200 kv. x-rays. The tyrosine concentration was decreased proportionally to the x-ray dose. Making the solutions acid by addition of H_2SO_4 did not sensibly change the results. The initial concentration of tyrosine did not greatly influence the absolute amount of tyrosine altered by a given dose. The authors concluded that "the sequence of the events, therefore, seems to be a primary ionization of water molecules and a following transference of energy from the ions to the amino-acid molecules. The phenol structure is evidently changed." They found the ratio, M_T/N, to be $^1/_6$ and $^1/_{12}$, respectively, for solutions of 1.0 mg. and 0.02 mg. of tyrosine per cm.^{3} M_T represents the number of tyrosine molecules altered by the rays and N is equal to the number of ion-pairs produced in the water. N was derived by assuming that the number of ion-pairs produced in the water was equal to the number produced in air by the absorption of x-rays equal to the quantity absorbed by the water.

About one year ago the author irradiated aqueous solutions (neutral) of tyrosine by α-rays. The experiments were performed in Stenstrom's laboratory. Since the experiments were not published, their essential features will be described briefly. Alpha-particles from radon were used. The technical procedure was similar to that used in other radiochemical studies.[6] The irradiated tyrosine was analyzed by three methods, but only the results of Folin and Ciocalteu's molybdate method[7] were selected for comparison since Stenstrom had used the same method in his x-ray experiments. In this manner, the effects of the decomposition products of the irradiated tyrosine in the molybdate reaction were equalized in the two sets of experiments.

The concentration of the solutions before irradiation was 0.04 mg. of tyrosine per cm.^{3} of water. After 13.7 mc. of radon were destroyed in 51 cm.^{3} of the solution 0.102 mg. of tyrosine were found to be decomposed. Calculation of M_TN gave $^1/_{350}$, where M_T has the same meaning as defined previously and N is the total number of ion-pairs produced in the water. N is derived from the product

$$E \cdot n \cdot (1.67 + 1.83 + 2.37)\ 10^5 \cdot i$$

where E is the quantity of radon (mc.) destroyed, n is the number of α-particles emitted by the destruction of 1 mc. of radon $(1.67 + 1.83 + 2.37) \cdot 10^5$ is the number of ion pairs produced in air by an α-particle from each of the three sources, *Rn*, *Ra A* and *Ra C*, and i is the total specific ionization of water. When referred to air as unity i is equal to 0.82.

The x-ray ($^1/_6$ to $^1/_{12}$) and α-ray ($^1/_{350}$) values of M_T/N are too widely separated to be accounted for by experimental errors. The disagreement suggests either that an ion causes chemical changes of various magnitudes, with respect to tyrosine in water, which depend on ionic concentrations, or that x-ray ions and α-ray ions produce physical and chemical changes in tyrosine which are not alike, and which cannot be analyzed correctly by the same method. The latter suggestion does not appear to be true. The solutions, either by x-ray or α-ray irradiation, became red-brown in color and their ultra-violet absorption spectra were changed exactly alike. However, the essential characteristics of the chemical and physical changes were not understood and the identification of the end products was not complete.

At least two secondary phenomena may influence the values of M_T/N. They are the neutralization of ions by recombination before chemical reaction occurs and the activation of molecules instead of ionization. Consider first the phenomenon of recombination. The determination of N for a given radiation dose is based on the measure of saturation currents in air. All of the ions are registered before they recombine. In solutions or in semi-solid biological media the same radiation dose will generate an approximately equal number of ions, but all of these ions may not react before recombination occurs. Therefore the calculated value of N may be greater than the actual number of ions which reacted chemically; consequently the calculated value of M_T/N is less than the true value. Recombination is thought to be greatest for the condition of highest ionic concentration; that is, in the "tracts" of protons and α-particles. For this reason, the α-ray action on tyrosine per calculated ion-pair should be smaller than the corresponding x-ray effect. This agrees with the experimental results.

Activation rather than complete ionization of molecules has been suggested as the mechanism of the primary radiation effects in liquids and in living cells.[8] For example, x-ray irradiated water is decomposed and hydrogen peroxide is formed through the primary activation of dissolved oxygen.[9] But activation is not the only effect on molecules in the liquid state, for carbon disulphide[10] and other dielectric liquids are ionized by x-rays. Moreover, α-rays decompose water. The primary process is ionization and the yield is one pair of ions for each molecule decomposed.

However, if activation by x-rays occurs and if the activated molecule can produce the chemical changes in the tyrosine molecules, then the tyrosine yield per ion-pair from x-ray irradiation should be proportionally larger. An activated molecule would not be registered as an ion in the determination of N. Thus, activation has the effect of increasing M_T without affecting N; whereas recombination reduces N without altering the measured value of M_T. Both phenomena and the conclusions therefrom agree with the results of the tyrosine experiments.

On the other hand, these phenomena are directly opposed to the "selective effects" of neutrons, for in Lawrences' experiments the x-ray values of M/N are smaller than the neutron values. Although tyrosine in water and living tissues are not analogous chemical systems, the major portion of the ionization in them is identical; that is, ionization of the water. Recombination and activation therefore play similar rôles in the two systems. But in tissues they are apparently masked by other effects if "selective effects" are real. Direct transformation of constituents of the cells other than water may attain magnitudes greater than the ionization theory postulates, and these transformations and their resultant secondary biological processes may determine the differential effects. In any case, investigators interested in the radiation effects in biological materials have chosen a working hypothesis which already has attained success in its explanation of radio-chemical reactions in gases and which promises success in its extension to more complex materials. But the extension of the ionization theory to radio-biological reactions is now restricted by want of a clear perception of the ionization processes in simple liquids and solids.

* The ionic concentration referred to is the number of ions in unit length of a "tract" and is not the number of "tracts" in unit volume of irradiated materials.

[1] John H. Lawrence and Ernest O. Lawrence, *Proc. Nat. Acad. Sci.*, **22**, 124 (1936).

[2] John H. Lawrence, Paul C. Aebersold and Ernest O. Lawrence, *Proc. Nat. Acad. Sci.*, **22**, 543 (1936).

[3] Raymond E. Zirkle and Paul C. Aebersold, *Proc. Nat. Acad. Sci.*, **22**, 134 (1936).

[4] Samuel C. Lond, *The Chemical Effects of Alpha Particles and Electrons* (1928).

[5] W. K. Stenstrom and A. Lohmann, *Jour. Biol. Chem.*, **79**, 673 (1928).

[6] Carl E. Nurnberger, *Jour. Chem. Phys.*, **4**, 697 (1936).

[7] Folin and Ciocalteu, *Jour. Biol. Chem.*, **73**, 627 (1927).

[8] Hugo Fricke, *Symposis on Quantitative Biology*, Biological Laboratory, Cold Spring Harbor, **2**, 241 (1934).

[9] H. Fricke and E. R. Brownscombe, *Phys. Rev.*, **44**, 240 (1933).

[10] F. L. Mohler and Lauriston S. Taylor, *Am. Jour. Roentg. Ra. Therapy*, **35**, 84 (1935).

EFFECTS OF THE LENGTH OF X-RAY WAVES ON SEEDS

By Arthur A. Bless

Department of Physics, University of Florida

Communicated March 8, 1937

The effects of x-rays on seeds have been investigated by a number of observers. It is generally agreed that a heavy dose of x-rays is injurious to the plants and retards growth. Some observers have noticed, however, that a small dose is in general beneficial to plants. Shull and Mitchell[1] have found that in the case of seeds of wheat and corn irradiated with x-rays, the plants have shown for a certain dose a marked increase in size and foliage. On the other hand Johnson[2] found no such increase for tomatoes, sunflowers and for other plants. The object of this investigation was to study the effects of moderate x-ray doses on seeds.

The experiments conducted by Shull were repeated on the type of seeds used by him, namely, the golden yellow dent corn. The experimental set-up was similar to that employed by Shull. An x-ray tube with a 45° target operated at 5 milliamperes and 100,000 volts was used. The seeds were placed 30 centimeters from the target and the only variable used was the time of exposure. The x-rays were filtered through a 1 mm. of aluminum.

The seeds after exposure to x-rays were planted in large pots in a greenhouse; there were 12 seeds in each pot. The height of the plants was measured at regular intervals and the average height was plotted as a function of the dose. For a certain date the plants of the seeds exposed for three minutes were largest in size, in agreement with the results obtained by Shull.

However, it was found that this increase over the other plants did not prevail for very long. When the growth of the plants is plotted for a later date the curve shifts and the maximum height shifts to another dose. At a later stage of the growth the plants of the seeds exposed for four minutes were larger than those of the three minute exposure. The optimum dose determined in this way evidently holds only for a given stage of the growth of the plant and cannot, therefore, be used as guide when the yield at maturity is desired. Several series of experiments have been made, which indicate the validity of this conclusion. This will be discussed later.

The biological effects of the wave-length of x-rays on matter seems to be a subject of controversy. From the point of view of the physicist the effects of x-rays are entirely due to the absorption of the radiation by the material, and the consequent ionization of the atoms produced directly by the x-rays as well as by the ejected photoelectrons. On this view there should be no difference between the action of long x-ray waves and short waves so long as the energy absorbed is the same. This view seems to be

held by a number of observers. On the other hand, some observers report that x-rays of long wave-lengths have specific effects, that they are injurious to life even in comparatively small doses.[3]

TABLE I

GROWTH CHART. AVERAGE HEIGHT IN CENTIMETERS

SERIES	TIME OF EXPOSURE IN MINUTES	WEEKS AFTER PLANTING								
		3	4	5	6	7	8	9	10	11
A	λ = 0.6A° 1	11.0	21.0	29.4	33.2	41.0	43.3	48.6	59.0	*
	2	10.8	20.0	29.7	36.1	47.0	52.0	58.3	69.3	
	3	10.8	21.0	30.9	37.0	44.6	49.0	57.0	66.3	
	4	10.6	17.7	27.7	33.8	46.0	53.6	60.6	74.3	
	5	13.6	24.5	34.4	40.5	53.0	60.3	75.3	87.3	
B	λ = 0.4 0	15.5	25.0	34.0	37.6	40.5	47.6	48.3	49.0	61.6
	½	16.4	26.7	36.8	41.3	44.4	53.6	54.0	55.6	64.0
	1	12.7	21.4	32.1	39.2	43.4	56.6	58.3	64.6	69.6
	2	14.0	24.5	34.7	41.0	45.0	50.3	52.3	54.3	61.0
	3	12.7	2C.0	29.6	35.5	40.0	50.6	52.3	56.6	66.3
	4	10.4	17.4	24.8	33.4	40.3	55.6	57.0	65.3	70.6
C	λ = 0.3 0	11.6	22.0	31.7	36.6	40.0	48.0	50.6	55.3	61.0
	½	10.4	20.3	28.3	34.4	38.4	48.6	51.3	56.6	66.3
	1	10.0	23.3	32.2	39.3	43.7	51.3	53.6	60.0	69.0
	2	14.1	27.6	37.0	42.5	47.6	53.6	55.6	60.6	65.6
	3	5.6	12.1	20.3	27.3	36.0	52.6	56.6	66.0	72.0
	4	5.4	10.6	14.2	22.4	33.8	50.0	52.3	61.0	65.3
	5	3.3	6.5	14.3	21.3	31.3	59.5	60.3	71.0	77.1
D	λ = 0.2 0	18.5	28.0	35.8	40.5	43.4	51.0	52.0	55.3	66.3
	½	16.1	26.6	33.7	39.3	43.0	51.3	52.6	56.3	66.0
	1	15.0	20.0	33.8	39.0	43.4	52.6	54.6	59.0	68.0
	2	14.8	23.4	32.2	38.9	41.9	52.0	53.6	57.3	63.3
	4	15.3	24.7	31.2	36.6	38.4	44.6	46.3	50.0	60.3
	6	4.2	8.1	15.8	24.8	35.3	57.0	59.6	65.3	70.0
E	λ = 0.12 0	11.1	21.8	30.9	35.8	44.6	47.0	54.3	66.0	*
	1	11.5	23.2	31.5	38.3	48.3	51.3	56.6	65.6	
	2	9.1	18.6	29.0	37.6	55.0	37.3	63.6	73.3	
	3	5.4	11.0	19.5	30.5	52.3	59.3	66.0	74.0	
	4	5.0	14.0	22.1	31.8	52.6	57.6	64.3	70.0	
	5	2.2	7.0	16.0	26.0	40.0	41.0	41.0	41.0	

* Plants too large to remain in pots.

In order to test the validity of either of these points of view exposure to x-rays of five different wave-lengths obtained by proper potential across the tube and proper filtering were made on seeds of corn. The effective

wave-length ranged from 0.6 to 0.12 angstrom unit. For each wave-length five or more exposures varying in time were made on different batches of seeds.

In order to have comparable results for all the series the distance of the seeds from the target was adjusted so that the seeds absorbed approximately the same quantity of x-ray energy for the same length of time of exposure. The seeds were planted in pots in a greenhouse. In all there were twenty-eight pots.

The growth of the plants was recorded every week and the results are tabulated below. An examination of the data discloses the following features:

(1) There seems to be no significant difference in the effects of radiation of the different wave-lengths used. The size of plants exposed to x-rays for an equivalent time of one minute, for example, at any given date is not much different for any given wave-length.

(2) Larger doses seem to exert inhibitive effects in most cases. (The five minute exposure of series one seems to be exceptional in this respect.) It is to be observed, however, that this inhibitive action is only temporary in character. The seeds in time attain the same height and in some cases a larger height inside of a few weeks. (The five minute exposure in series 5 should be disregarded. Something was obviously wrong with this particular set as the seeds have not shown any growth for several weeks.)

(3) It was mentioned previously that the optimum dose at one date may be different from the optimum dose at another date. For example, in series C seven weeks after planting the two minute dose seems to produce a greater height of plants than any other dose. However, a week later the five minute exposure seems to yield best results.

Another series of experiments with 21 pots was made very recently giving essentially the same results.

I wish to thank Dr. F. H. Hull and Dr. R. W. Rupprecht of the Agricultural Experiment Station for their kind assistance.

[1] C. A. Shull and J. A. Mitchell, *Plant Physiol.*, **8**, 287 (1933).

[2] E. A. Johnson, *Am. Jour. Bot.*, **18**, 603 (1931).

[3] A comprehensive review of the literature on this subject is given in "Biological Effect of Radiation," Vol. 1, Ch. XIII, McGraw-Hill Book Co.

SYMMETRY OF THE TRANSMISSION COEFFICIENTS FOR THE PASSAGE OF PARTICLES THROUGH POTENTIAL BARRIERS

BY MORRIS MUSKAT AND ELMER HUTCHISSON

GULF RESEARCH & DEVELOPMENT COMPANY AND UNIVERSITY OF PITTSBURGH

Communicated February 24, 1937

I. *Introduction.*—In problems involving the transmission of electrons through potential barriers, such as occur in the theory of contact rectification,[1] it is usually stated or assumed that the transmission coefficient is independent of the direction of motion of the incident electron. Referring to figure 1, such a statement means that the transmission coefficient for electrons coming from the left is the same as that for electrons of the same total energy incident on the barrier from the right.

Although this result may be shown to follow from an application of the principle of detailed balancing to an assembly of particles in thermodynamic equilibrium in a closed system, no general wave mechanical proof appears to have been given. In view of its elementary nature and generality, it seems desirable to present the following proof of the symmetry of the transmission coefficient even though it may be verified in the few special cases already in the literature.

II. *Non-Relativistic Case.*—It is convenient to consider the potential energy function $U(x)$ as divided into three regions, as indicated in figure 1. No assumption is made as to the nature of the variations of U at the boundaries of the regions.[2] Of the two independent solutions of the wave equation:

$$\frac{d^2\Psi}{dx^2} + \frac{8\pi^2 m}{h^2}\left\{ W - U(x) \right\}\Psi = 0, \tag{1}$$

in each region, we shall denote by v that traveling toward the left and by u that traveling toward the right. Subscripts will be used to denote the regions to which the wave functions refer. The functions in regions 1 and 3 will be taken as normalized.

If the wave is considered as incident from the right the boundary conditions of the continuity of Ψ and its first derivative require:

$$\left.\begin{aligned} A_2v_2(a) + B_2u_2(a) &= A_3v_3(a) + B_3u_3(a), \\ A_2v'_2(a) + B_2u'_2(a) &= A_3v'_3(a) + B_3u'_3(a), \\ A_1v_1(0) &= A_2v_2(0) + B_2u_2(0), \\ A_1v'_1(0) &= A_2v'_2(0) + B_2u'_2(0), \end{aligned}\right\} \tag{2}$$

where the A's and the B's are the amplitudes of the wave functions and primes indicate derivatives with respect to x. Denoting by α_1 and α_3

the velocities in the regions 1 and 3, the transmission coefficient may be written:

$$D_1 = \frac{|A_1|^2}{|A_3|^2}\frac{\alpha_1}{\alpha_3}. \tag{3}$$

Solving equation (2), we obtain:

$$D_1 = \frac{\alpha_1}{\alpha_3}\frac{|C_2(0)|^2|C_3(a)|^2}{|\Delta|^2}, \tag{4}$$

where $C_j(x)$ is the Wronskian of u, v and Δ is a determinant which will cancel in the final result. Similarly if the wave is incident from the left:

$$D_2 = \frac{\alpha_3}{\alpha_1}\frac{|C_1(0)|^2|C_2(a)|^2}{|\Delta|^2}. \tag{5}$$

Noting that the Wronskian $C_j(x)$ is independent of x since equation (1) does not contain a first derivative,[3] it follows that $C_2(0) = C_2(a)$.

Assuming u_j and v_j to have the form: $e^{\mp i\kappa_j x}$ in the regions 1 and 3 where $\kappa_j^2 = 8\pi^2 m(W - U_j)/h^2 = 4\pi^2 m^2 \alpha_j^2/h^2$ we have from equation (4):

$$D_1 = \frac{16\pi^2 m^2 |C_2|^2 \alpha_1\alpha_3}{h^2|\Delta|^2}. \tag{6}$$

As the same expression is obtained from equation (5), we have the result that the transmission coefficient D_1 for waves from the right is equal to D_2 for waves from the left.

The above result may be generalized to the case in which there are a finite number of discontinuities in region 2. If these be at $0, a_1 \ldots a_n$, where 0 and a_n enclose the region 2, equations (4) and (5) become:

$$D_1 = \frac{\alpha_1}{\alpha_{n+2}}\frac{|C_2(0)|^2|C_3(a_1)|^2\ldots|C_{n+2}(a_n)|^2}{|\Delta|^2}, \tag{7}$$

and

$$D_2 = \frac{\alpha_{n+2}}{\alpha_1}\frac{|C_1(0)|^2|C_2(a_1)|^2\ldots|C_{n+1}(a_n)|^2}{|\Delta|^2}. \tag{8}$$

As before, it follows at once that $D_1 = D_2$.

The same result may be obtained if in place of the assumption that Ψ and Ψ' are continuous at the boundaries of the regions, one requires the continuity of $\beta\Psi$ and $\gamma\Psi'$ where β and γ are constants characteristic of the various regions. Under these conditions, equations (4) and (5) become:

$$D_1 = \frac{\alpha_1}{\alpha_3} \frac{\beta_2^2\beta_3^2\gamma_2^2\gamma_3^2 \mid C_2(0) \mid^2 \mid C_3(a) \mid^2}{\mid \Delta \mid^2}, \tag{9}$$

and

$$D_2 = \frac{\alpha_3}{\alpha_1} \frac{\beta_1^2\beta_2^2\gamma_1^2\gamma_2^2 \mid C_1(0) \mid^2 \mid C_2(a) \mid^2}{\mid \Delta \mid^2}. \tag{10}$$

The result that $D_1 = D_2$ therefore depends only upon the equality:

$$\beta_1\gamma_1 = \beta_3\gamma_3, \tag{11}$$

which is less restrictive than the original assumption where $\beta_1 = \beta_2 = \beta_3$ and $\gamma_1 = \gamma_2 = \gamma_3$.[4]

III. *Relativistic Case.*—With the notation:

$\Phi_1 = \Psi_1 + \Psi_4$; $\Phi_2 = \Psi_2 + \Psi_3$; $\Phi_3 = \Psi_2 - \Psi_3$; $\Phi_4 = \Psi_1 - \Psi_4$

$$p = \{W - U(x)\}/c; \quad \epsilon = h/2\pi i, \quad \} \tag{12}$$

the second order one-dimensional wave equations for the Dirac electron may be written as:

$$\left.\begin{aligned} \epsilon\Phi''_{1,2} + \left(\frac{m^2c^2}{\epsilon} - \frac{p^2}{\epsilon} + p'\right)\Phi_{1,2} = 0, \\ \epsilon\Phi''_{3,4} + \left(\frac{m^2c^2}{\epsilon} - \frac{p^2}{\epsilon} - p'\right)\Phi_{3,4} = 0. \end{aligned}\right\} \tag{13}$$

with:

$$p\Phi_{1,2} + mc\Phi_{4,3} + \epsilon\Phi'_{1,2} = 0.$$

Since Φ_1 and Φ_2 and Φ_3 and Φ_4 satisfy the same equation, we shall hereafter take $\Phi_1 = \Phi_2$ and $\Phi_3 = \Phi_4$.

Setting further:

$$\left.\begin{aligned} \Phi_{kj} &= Av_{kj}(x) + B_{kj}u_{kj}(x), \\ \mu_{kj}(x) &= \frac{1}{mc}[p(x)u_{kj}(x) + \epsilon u'_{kj}(x)], \\ \nu_{kj}(x) &= \frac{1}{mc}[p(x)v_{kj}(x) + \epsilon v'_{kj}(x)], \end{aligned}\right\} \tag{14}$$

where $k = 1, 2, 3$ or 4 and refers to the Dirac wave function, and $j = 1, 2$ or 3 and refers to the potential region, it follows that:

$$\left.\begin{aligned} B_{1j}\mu_{1j}(x) &= -B_{4j}u_{4j}(x), \\ A_{1j}\nu_{1j}(x) &= -A_{4j}v_{4j}(x). \end{aligned}\right\} \tag{15}$$

The condition of continuity of Φ at $x = a$ and $x = 0$ may then be shown to give for incident waves from the right:

$$\frac{A_{11}}{A_{13}} = \frac{A_{21}}{A_{23}} = \frac{\epsilon^2 C_{12}(0) C_{13}(a)}{m^2 c^2 \Delta}, \tag{16}$$

where Δ is again a determinant which will cancel out in the final result. For waves incident from the left we find similarly:

$$\frac{B_{13}}{B_{11}} = \frac{B_{23}}{B_{21}} = \frac{\epsilon^2 C_{11}(0) C_{12}(a)}{m^2 c^2 \Delta}. \tag{17}$$

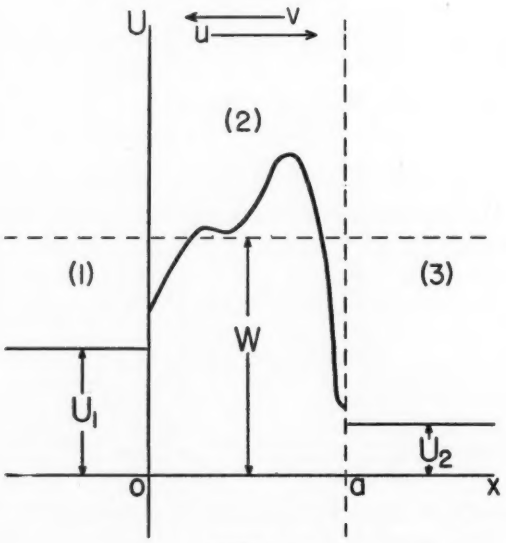

FIGURE 1

The current density may be expressed as:

$$s = \frac{c}{2} [\Phi_1 \Phi_1^* - \Phi_4 \Phi_4^* + \Phi_2 \Phi_2^* - \Phi_3 \Phi_3^*]. \tag{18}$$

The incident current for waves from the right is, therefore:

$$s_i = c \mid A_{13} \mid^2 \{ \mid v_{13}(a) \mid^2 - \mid \nu_{13}(a) \mid^2 \}. \tag{19}$$

Likewise the transmitted current is:

$$s_t = c \mid A_{11} \mid^2 \{ \mid v_{11}(0) \mid^2 - \mid \nu_{11}(0) \mid^2 \}. \tag{20}$$

Hence the transmission coefficient is:

$$D_1 = \frac{s_t}{s_i} = \frac{\mid A_{11} \mid^2}{\mid A_{13} \mid^2} \left\{ \frac{\mid v_{11}(0) \mid^2 - \mid \nu_{11}(0) \mid^2}{\mid v_{13}(a) \mid^2 - \mid \nu_{13}(a) \mid^2} \right\}. \tag{21}$$

Similarly, for waves from the left, the transmission coefficient may be shown to be:

$$D_2 = \frac{|B_{13}|^2}{|B_{11}|^2}\left\{\frac{|u_{13}(a)|^2 - |\mu_{13}(a)|^2}{|u_{11}(0)|^2 - |\mu_{11}(0)|^2}\right\}. \tag{22}$$

Taking, as before, solutions of the form:

$$u_{kj} = e^{-i\kappa_j z} \text{ and } v_{kj} = e^{i\kappa_j z} \text{ for } j = 1 \text{ and } 3,$$

we obtain, on applying equations (16) and (17) and noting again that the Wronskians are independent of x:

$$D_1 = \frac{4\epsilon^4 |C_{12}|^2 \kappa_1\kappa_3}{m^4c^4|\Delta|^2}\left(\frac{p_1 + i\epsilon\kappa_1}{p_3 + i\epsilon\kappa_3}\right), \tag{23}$$

where p_1 and p_3 refer to the value of p in the regions 1 and 3. Further:

$$D_2 = \frac{4\epsilon^4 |C_{12}|^2 \kappa_1\kappa_3}{m^4c^4|\Delta|^2}\left(\frac{p_3 - i\epsilon\kappa_3}{p_1 - i\epsilon\kappa_1}\right), \tag{24}$$

which is evidently the same as D_1.

The authors desire to thank Professor A. E. Ruark for the opportunity of discussing this problem with him.

[1] J. Frenkel, *Phys. Rev.*, **36**, 1604 (1930); L. Nordheim, *Zeits. Phys.*, **46**, 833 (1928).

[2] In fact, 0 and a simply represent the division points beyond which (i.e., <0, and $>a$) the potential may be considered as constant. Except in the cases where $U(x)$ oscillates or becomes infinite at $x = \pm\infty$, $U(x)$ will always approach constancy, at least asymptotically, for x sufficiently large.

[3] Cf. E. L. Ince, *Ordinary Differential Equations* (1927), p. 119.

[4] In any one-dimensional wave problem the restrictions on the boundary conditions may be derived in a similar manner by requiring the symmetry of the transmission coefficient of the appropriate physical quantity (such as the energy flux in the case of sound waves).

POLYMORPHIC TRANSITIONS OF INORGANIC COMPOUNDS TO 50,000 kg./cm.²

BY P. W. BRIDGMAN

THE PHYSICS LABORATORY, HARVARD UNIVERSITY

Communicated February 23, 1937

In the following I present a summary of results which will be published in full detail, probably after some delay, in *Proceedings American Academy of Arts and Sciences*. Methods recently devised[1] by which the range of pressures open to experiment is increased to 50,000 kg./cm.² are applied to a study of polymorphism in a number of inorganic compounds. Thirty-five such compounds have been found to exhibit polymorphism in this pressure range, and the transition parameters (pressure, temperature, change of volume and latent heat) for these transitions have been determined. In addition, more than 50 other compounds have been examined for polymorphism with negative results.

Rough average values of the parameters of the 20 of these substances which are best adapted to this sort of summary are collected in the table. The temperature range of the measurements was from −79°C (solid CO_2 temperature) to 200°; the average temperature referred to in the table was about 100°C. In the detailed paper will be found complete diagrams and numerical tables giving the various parameters as a function of pressure and temperature.

The 15 substances found to exhibit high pressure polymorphism not given in this summary are: AgCl, AgBr, HgI_2, GeI_2, PbI_2, Cr_2O_3, AgCN, $NaNO_2$, $CsNO_3$, $NaIO_4$, $KClO_3$, $CsIO_4$, NH_4ClO_4, Na_2SO_4 and $KMnO_4$. For some of these the transitions are sluggish or obscure, while for others the variation of the parameters is so great as to make it undesirable to summarize with a single number. Of these PbI_2 has two transition lines (3 modifications) and Na_2SO_4 two lines (three modifications). Several other modifications of Na_2SO_4 are known at atmospheric pressure at higher temperatures. Each of the others in this list of 15 substances has only one new high pressure modification.

The number of examples of high pressure polymorphism presented here is sufficient to justify a statistical study. The most obvious result is that the rôle of temperature does not become markedly different in this new high pressure range. If the results of my previous measurements[2] of polymorphism up to 12,000 kg./cm.² are plotted on the same scale as the new results, the general aspect of the diagrams is the same, almost all the transition lines sweeping through the entire temperature range of 200° for a pressure change of 10,000 kg./cm.² or less. Since the third law demands that the transition line be vertical at 0°Abs., this means that

most of the transition lines will plunge into the pressure axis at 0°Abs., so that polymorphism must be a common phenomenon at 0°Abs. There is no apparent tendency for the pressures of transition to crowd toward either high or low values, so that statistically polymorphism at 0°Abs. is uniformly distributed over the pressure range.

TABLE 1

SUBSTANCE	MEAN PRESSURE KG./CM.²	MEAN $\frac{d\tau}{dp}$	MEAN VOLUME CHANGE, PER CENT	MEAN LATENT HEAT KG. CM./GM.
$HgCl_2$	16,500	−0.018	0.7	29.0
$HgBr_2$	2,200	0.167	1.1	4.5
	24,000	0.046	0.6	8.4
	38,500	−0.094	1.3	8.4
$AgNO_2$	22,500	0.056	2.7	37.0
KCN	4,400	−0.088	2.9	85.0
	21,200	+0.090	8.0	220.0
	19,000	0.042	9.4	360.0
	32,200	0.005	0.4	197.0
$NaClO_3$	15,000	−0.046	0.1	2.0
	32,000	−0.016	1.1	100.0
	28,000	−0.010	1.5	250.0
	36,000	0.008	3.0	63.0
$NaBrO_3$	18,900	−0.027	2.5	116.0
$NaClO_4$	25,300	0.017	0.4	30.0
	32,500	+0.018 −0.013	3.7	0 ±
KIO_4	12,800	0.074	0.4	4.6
$CsClO_4$	1,700	0.114	4.9	48.0
$TlClO_4$	11,900	−0.125	2.2	13.0
$RbClO_4$	23,000	−0.024	1.7	100.0
$AgClO_4$	8,300	0.118	6.4	42.0
$AgIO_3$	14,600	−0.071	0.8	8.0
Cu_2I_2	14,500	−0.060	3.5	30.0
Ag_2S	19,000	0.014	0.04	16.0
$CsMnO_4$	12,900	−0.030	2.2	75.0
	39,500	−0.048	4.9	106.0
$ZnBr_2$	29,000	0.050	7.4	125.0
$AgNO_3$	35,000	0.033	1.6	42.0
	40,200	−0.033	4.9	125.0
$Pb(C_2H_3O_2)_2$	2,500	0.026	2.0	100.0
	2,500	0.030	1.1	60.0
	12,200	0.029	0.4	170.0
$RbNO_3$	21,300	0.021	0.8	45.0

A study in finer detail of the variation in the mean range of roughly 200° for 10,000 kg./cm.² does disclose a significant trend at high pressures. In the previous work up to 12,000 kg./cm.², 14 out of the 59 transition lines examined, or 24 per cent, were of the ice type, that is, the phase stable at

the higher temperature has the smaller volume. One would be inclined to regard this as abnormal, and expect it to become less frequent at high pressures. In support of this expectation there is the fact that the only three known cases of an ice type of transition between liquid and solid, namely water, bismuth and gallium, are characteristic only of low pressures, at higher pressures there being new polymorphic modifications with the normal volume relation between solid and liquid. This expectation turns out to be incorrect, however, for of 44 transition lines measured in this new work up to 50,000 kg./cm.^{2}, 19, or 43 per cent, are of the ice type. The ice type of transition between solids therefore tends to become more common at high pressures, unlike the ice type of transition between solid and liquid.

A statistical study may be made of the volume changes of these transitions by plotting volume change against the mean pressure of the transition. It will be found that there is no significant change with pressure, comparatively large volume changes being just as likely to occur at high pressures as at low. This is perhaps not what might be expected. The volume change may also be studied statistically in another way by plotting the number of examples against the logarithm of the percentage volume change. In this way a distribution curve will be found of roughly constant amplitude from a volume change of 11 per cent down to 0.8 per cent. Below this, the distribution amplitude tails off, eventually disappearing at the smallest volume change measured, 0.04 per cent. An adequate explanation of this tailing off is the experimental difficulty of measuring small volume changes, so that as far as the present results go, there is nothing to indicate that the volume changes of polymorphic transition are not distributed with an approach to uniformity over the entire logarithmic scale from 11 per cent down, or that pressure polymorphism will not be found to be an increasingly common phenomenon as the sensitiveness of the measurements is increased.

A somewhat similar statistical situation is presented by an analysis of the number of high pressure phases presented by different substances. A summary of all my results, both my former ones to 12,000 and the present ones to 50,000 kg./cm.^{2} yields the following:

Number of phases	2	3	4	5	6	7
Number of examples	45	13	7	0	3	1

Very roughly, and with due regard for the small number of examples, one may say that the number of examples falls off geometrically as the number of phases increases arithmetically, which means that the chance that a substance will exhibit polymorphism with increasing pressure is independent of whether it has already exhibited polymorphism at some lower pressure.

The latent heat of the transition will be found to exhibit no statistical trend with increasing pressure. The energy change during the transition, $\Delta E = L - p\Delta v$, on the other hand tends statistically to increase proportionally to the pressure, since neither L nor Δv have any statistical trend with pressure. This means that on the average the thermal effect becomes continually relatively less important, and that the mechanical effect, expressed by the $p\Delta v$ term, comes to dominate. Under familiar conditions at low pressures the heat usually dominates the situation, so that for a transition of the normal type (high temperature phase having the larger volume) the high pressure modification has a smaller energy content than the low pressure phase, the energy which flows out as latent heat when the transition runs not being compensated by the inflowing mechanical work. Among the transitions examined here, however, the latent heat overbalances the mechanical work in only 22 per cent of the cases, so that the internal energy of the high pressure modification is greater in 78 per cent of the cases. This is strikingly different from the situation for melting, since in all known cases, and up to 12,000 kg./cm.[2], the latent heat overbalances the mechanical work by a large amount.

Schottky's theorem gives the change of the total internal kinetic energy under our conditions: $\Delta E_{kin} = 4p\Delta v - L$. The latent heat, L, is so far overbalanced by the term $4p\Delta v$ that for this rough discussion we may disregard L and put $\Delta E_{kin} = 4p\Delta v$. Since Δv is necessarily negative, it follows that the total internal kinetic energy decreases on passing from the low to the high pressure phase. Part of ΔE_{kin} arises from the change in the zero point energy, which we may plausibly expect to increase on passing to the high pressure modification. The part of ΔE_{kin} arising from the internal motions of the electrons in the atoms and molecules must therefore decrease by more than $4p\Delta v$. If the atoms in the solid preserve anything approaching to the same over-all electronic structure that they have in the isolated condition, this must mean an *increase* in the size of the atom on passing to the high pressure form. This result is so highly paradoxical that one seems almost driven to the conclusion that an important fraction of the electrons are in an essentially different state from that in the free atoms, and that polymorphic transition involves an important change in the state of these electrons. It suggests itself, therefore, that the clue to the explanation of these polymorphic changes, which up to now has so obstinately resisted theoretical attack, is to be found in something similar to the coöperative phenomena between the electrons of the entire structure which are already known to be determinative for electrical resistance.

[1] Bridgman, P. W., *Phys. Rev.*, New York, **48**, 893–906 (1935).

[2] See, for example, the summary in my book, *The Physics of High Pressure*, Bell (London) or Macmillan (New York), 1931.

COLOR CHANGES DUE TO ERYTHROPHORES IN THE SQUIRREL FISH HOLOCENTRUS*

BY G. H. PARKER

BERMUDA BIOLOGICAL STATION AND HARVARD BIOLOGICAL LABORATORIES

Communicated March 11, 1937

Fishes in which the most rapid color changes occur are those that are predominately red and whose integumentary color-cells are the well-known erythrophores. A good example of this kind is Holocentrus ascensionis (Osbeck) of Bermuda. This fish, according to Smith and Smith (1935), will change red to white in some five seconds and white to red in about ten seconds. In my own tests individuals of this species changed from red to white in from 4.5 to 8.5 seconds with an average from thirty trials on sixteen fishes of 6.38 seconds and from white to red in from 16.5 to 22.0 seconds with an average of 18.7 seconds (19°C). These changes were induced by gently transferring the fish from a black-walled illuminated bowl to a white-walled one or the reverse. Whether the blanching was in part excitement pallor, as might be inferred from the observation of Abramowitz (1936), was not determined. This change seemed at times almost instantaneous as von Frisch (1912) found the excitement pallor in Trigla to be. These rapid changes are in strong contrast with the erythrophore responses of Scorpaena which require hours or even a day (Smith and Smith, 1934) for completion. Are the changes in Holocentrus due to nervous or to humoral factors or to both?

When a deep cut is made parallel to the lateral line and dorsal to it on a white Holocentrus, a red band appears almost at once extending from the cut dorsally and posteriorly over the trunk of the fish and into its dorsal fin. This band is due to the dispersion of the red pigment in the erythrophores which in appearance change from minute red dots to roundish, irregular, red splotches with about four times the diameter of the dots (Fig. 1). If a similar flesh cut is made ventral to the lateral line, a less obvious red band will form from this cut ventrally. Like conditions have been noted by Smith and Smith (1934) in Scorpaena. In my experience the red bands in Holocentrus become faint in a few days. Smith and Smith (1935) record large red areas due to nerve cutting in this fish as lasting ten to fifteen days. Flesh cuts in red fishes are not productive of visible bands, but when such a red fish is caused to blanch in a white bowl, the red band will make its appearance in appropriate position and form showing that it was merely hidden in the general redness. The concentrated and the dispersed conditions of the erythrophore pigment in these changes have been well photographed by Smith and Smith (1935). Beside erythrophores,

Holocentrus possesses many xanthophores, and some melanophores. The present studies deal with the erythrophores only.

It has been customary since the time of Brücke (1852) to interpret the dispersed condition of chromatophores resulting from cutting their nerves as due to nerve paralysis with the consequent lapse of the color cells into a state of rest. But I have shown (1934, 1936) that the cutting of melanophore nerves throws these nerves into a state of temporary excessive activity that may last hours or even days, for after this activity has subsided it can be revived by recutting the nerves. Such reactivation would be quite impossible were the state of the cut nerves one of paralysis. This phenomenon, demonstrated on melanophores, can also be shown on erythrophores. The best place in Holocentrus to test this matter is the dorsal fin.

The dorsal fin of Holocentrus consists of about eleven erectile, sharp rays rather regularly spaced and connected one with another by a delicate transparent membrane. Behind the eleventh ray the fin is continued in a series of crowded much less regular rays that reach almost to the root of the tail. The membrane that unites one ray with another in the anterior regular part of the fin is dotted over with erythrophores. When tautly expanded by the erection of the rays, this membrane is roughly rectangular in outline and in the deepest part of the fin in a fish about 18 cm. long measures 15 mm. from the fish's body to the free edge of the fin and 5 mm. from ray to ray. Such a membrane presents a most convenient object for experimentation on erythrophores (Fig. 1).

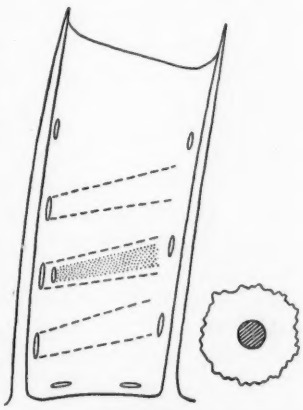

FIGURE 1

Side view of one segment in the dorsal fin of Holocentrus. The ray to the left is the anterior one. Cuts are shown through the interray membrane near its edges. Those next to the posterior and the ventral edges were not followed by the formation of red bands. Those next to the lower part of the anterior edge showed primary bands which, after fading, could be revived by a new cut as indicated in the middle one of the three. The side figure is the circular outline of the concentrated pigment in an erythrophore superimposed on the area of its dispersed pigment. The latter has about four times the diameter of the former.

As already stated when a deep flesh wound is made in a white Holocentrus between its lateral line and its dorsal fin, not only does a red band form dorsally and a little posterior to the wound over the trunk, but this band extends over the lower three-quarters of the membrane of the dorsal fin. As can be seen under the microscope, the erythrophores in this part of the

fin, at first punctate, are, after the cut has been made, fully expanded and the membrane in consequence changes from almost colorless to red. These erythrophores must, therefore, be innervated. What courses do the nerves take from the trunk of the fish to the color cells in the membrane? This can be determined by local cutting. If the membrane is punctured close to its regions of attachment by a narrow sharp knife blade, no color change is noticed on its surface so long as the punctures lie next the posterior fin-ray or next the line of the trunk, but when they lie next the proximal three-quarters of the anterior ray, every cut is followed by a band of red which, starting at the cut itself, passes diagonally in a postero-dorsal direction completely across the membrane almost to the posterior ray (Fig. 1). Such delicate bands are most sharply and clearly seen and show that the innervation of each interray membrane is from the posterior edge of the anterior ray and not from other parts.

These bands under the microscope can be shown to be made up of erythrophores with dispersed pigment as contrasted with those in the adjacent, more nearly transparent region which are dot-like. Such bands in white fishes gradually disappear in from four to twenty-four hours or more. In red fishes they remain as long as the fishes can be kept alive in aquaria, which in my experience is some six days or so. When such a red fish is temporarily blanched, the band can be clearly seen, but it is lost to view again when the fish reverts to red. These conditions are precisely parallel to what is known of the dark melanophore bands produced by cutting nerves in the tail of Fundulus.

If, after a red band in a fin membrane has faded, a new cut is made posterior to the initial cut and within the area of the band, a new band will commonly appear and cover the old band from the new cut to the posterior termination of that band much as Smith and Smith (1935) have demonstrated for the body at large. In three of my fishes in which what may be called the primary bands had been excited forty hours previously and had faded, three new cuts were made in the bases of the old bands with the result that in two instances new clear secondary bands were formed while in one no such band appeared. This record shows that forty hours after a nerve has been cut, it may still be reactivated, though in some instances this may fail. What the full-time limit for reactivation of such nerves is was not determined. Perhaps forty hours after the initial cut is not far from it. From these results I conclude that the erythrophores of Holocentrus, like the melanophores of Fundulus and other teleosts, are provided with dispersing nerve-fibres.

Concentrating nerve-fibres in the chromatophore systems of fishes has been commonly sought by electric stimulation. If the medulla of a Holocentrus is exposed by a dorsal incision at the hind end of the skull, the whole fish as a result of the operation will become red. On stimulating the medulla

by an electric current from an inductorium, the red fish almost immediately blanches, as was shown by Smith and Smith on Scorpaena (1934) and later on Holocentrus (1935). On cessation of stimulation Holocentrus quickly reddens. This test can be repeated many times over the half hour following the exposure of the medulla. If, prior to such a test, a considerable flesh cut is made dorsal to the lateral line of the fish, a large red area may be established as an area of reference. If now the medulla is stimulated electrically, the whole fish whitens except the denervated area which remains red and conspicuously so in contrast with the whiteness of the rest of the fish. Red fishes with exposed medullas blanch on electric stimulation in from five to seven seconds and return to red in from eight to twelve seconds. I conclude from these observations that the erythrophores of Holocentrus possess concentrating nerve-fibres as well as dispersing ones and that in this respect they are also like the melanophores of Fundulus. These erythrophores must, therefore, be doubly innervated. Of their two sets of fibres the dispersing fibres must have a low threshold for stimulating by cutting and probably a high one for electric stimulation and the concentrating fibres a low one for electric stimulation and probably a high one for cutting. Here again there is general agreement between the erythrophore system of Holocentrus and the melanophore system of Fundulus.

What part do the secretion of the pituitary gland and such neurohumors as adrenalin play in the color changes of Holocentrus? Four red fishes which had been cut for red bands and which from test were known to have developed these bands were injected each with 0.1 cc. adrenalin 1:1000 (Parke, Davis and Company) and were then put in black-walled bowls for observation. The four fishes in three minutes were fully pale, and paler than pale fishes rendered so normally except for the bands which remained red. These bands, however, soon began to fade and in about two hours they too were entirely pale, the part of the band on the body of the fish blanching before that on the dorsal fin. These tests were repeated and in all instances Holocentrus was found to blanch to strong adrenalin.

Three pale fishes were injected each with 0.1 cc. obstetrical pituitrin (Parke, Davis and Company) and placed in white-walled bowls for inspection. In ten minutes the region of injection in each was the center of a large red area which remained evident for four hours after which it faded. In no instance did the whole fish become red, but as microscopic inspection showed the red area in the region of injection was caused by the dispersion of the erythrophore pigment. An injection of pituitrin had no visible effect upon a red fish nor were injections of Ringer's solution followed by any color changes either in white fishes or in red fishes. These observations show that the erythrophores of Holocentrus are responsive to such neurohumors as adrenalin and pituitrin. Are there any such hormones normally active in the body of this fish?

To test this question blood transfers were resorted to, for such hormones are commonly carried in the blood. A number of squirrel fishes were kept some three days in illuminated white-walled vessels to induce a considerable formation of a blanching hormone if such there were. These fishes were then quickly bled, their blood defibrinated and injected part into red fishes and part into white ones. In none of these was there any change of color either local or general. Corresponding tests in which the defibrinated blood from red fishes was injected into white and into red individuals had the same result. In none of these transfers was there any evidence of a blood-borne hormone either for the dispersion or for the concentration of pigment.

As a further test of this matter the pituitary glands were taken from four squirrel fishes, minced in a small amount of Ringer's solution and 0.1 cc. of the extract thus obtained was injected into a white fish and the same amount into a dark fish. In neither instance was there in the course of three hours any sign of color change. From this and the preceding tests I am led to conclude that in the erythrophore color changes of Holocentrus no water-soluble hormone, hydrohumor, is involved and that these changes, as already maintained by Smith and Smith (1935), are under nervous control exclusively. This is in strong contrast with the condition in Phoxinus where the erythrophore changes, according to Giersberg (1930), are effected by the hypophysis. I believe the evidence presented in this paper justifies the further conclusion that the erythrophores in Holocentrus have a double innervation consisting of dispersing and of concentrating nerve-fibres. From the way in which red bands disappear I infer that the concentrating fibres act on the red color cells through a neurohumor soluble in lipoids, a lipohumor. This is probably also true of the dispersing fibres.

* For the facilities to carry out this work I am greatly indebted to the Bermuda Biological Station and its Director, Dr. J. F. G. Wheeler. The expenses of the investigation were covered by a generous grant from the Milton Fund of Harvard University, to the officers of which I express my sincere thanks.

A. A. Abramowitz, "Physiology of the Melanophore System in the Catfish, Ameiurus," *Biol. Bull.*, **71**, 259–281 (1936).

E. Brücke, "Untersuchungen über den Farbenwechsel des afrikanischen Chamäleons," *Denkschr. Akad. Wiss. Wien, math.-nat. Cl.*, **4**, 179–210 (1852).

H. Giersberg, "Der Farbwechsel der Fische," *Zeit. vergl. Physiol.*, **13**, 258–279 (1930).

K. von Frisch, "Ueber farbige Anpassung bei Fischen," *Zool. Jahrb., all. Zool. Physiol.*, **32**, 171–230 (1912).

G. H. Parker, "The Prolonged Activity of Momentarily Stimulated Nerves," *Proc. Nat. Acad. Sci.*, **20**, 306–310 (1934).

G. H. Parker, *Color Changes in Animals in Relation to Nervous Activity*, Philadelphia, 74 pp. (1936).

D. C. Smith and M. T. Smith, "Observations on the Erythrophores of Scorpaena Ustulata," *Biol. Bull.*, **67**, 45–58 (1934).

D. C. Smith and M. T. Smith, "Observations on the Color Changes and Isolated Scale Erythrophores of the Squirrel Fish, Holocentrus ascensionis (Osbeck)," *Biol. Bull.*, **68**, 131–139 (1935).

SOME QUANTITATIVE RELATIONS BETWEEN VISUAL STIMULI AND THE PRODUCTION OR DESTRUCTION OF MELANIN IN FISHES

BY F. B. SUMNER AND P. DOUDOROFF

SCRIPPS INSTITUTION OF OCEANOGRAPHY, UNIVERSITY OF CALIFORNIA

Communicated February 16, 1937

The ability of certain animals to conform their general color-tones to those of their backgrounds raises various questions of high theoretical interest. Some of these, curiously enough, are rather closely related to certain problems of human psychology.

One of these last is the so-called problem of "color constancy" which has received considerable attention from experimental psychologists. By this term is meant the seeming constancy of our perceptions of a given color or shade under widely different conditions of illumination.

Another of these problems is the relation between the intensity of a stimulus and the intensity of the resulting sensation. Here fall such phenomena as those for which the "Weber-Fechner law," was formulated.

To consider the first of these problems, how do we recognize a neutral gray surface as being gray rather than white, even in the absence of a white surface with which we may compare it? The answer seems plain. No single object in the visual field is an isolated fact; it is influenced more or less by the entire field. We make allowance, subconsciously at least, for the degree of illumination to which this field is subjected. When the conditions are such as to deceive the observer concerning the source of illumination, he may fall into serious error regarding the color or shade of the surface which he is looking at. Some earlier experiments seem to show that a fish may be similarly "deceived," so as to react in a manner which would be appropriate to a lighter or darker background than the one actually present.[1]

We know nothing of a fish's perceptions, but we do know considerable regarding its responses to various optic stimuli. It has long been known that the adjustments of a fish to a given background are largely independent of the intensity of the incident light. A ten-fold increase or decrease in the illumination of this background is likely to cause little or no change

in the shade assumed by the fish. It was early pointed out that the adaptive value of these color changes, in their relation to concealment, requires that this should be the case.[2]

It seems plain, then, that the chromatic reactions of a fish are determined neither by the intensity of the incident light nor by the intensity of the light reflected from the background, but by some relation between these two factors. Which means, otherwise stated, by some relation between different parts of the visual field. The suggestion is somewhat obvious[3] that the ratio between the light reflected from the nearby surfaces within the experimental tank and the light which enters the tank from above constitutes the effective stimulus in this matter.[4] For this ratio we have a convenient term, albedo. A gray, for example, has a lower albedo than white, though it is plain that under proper relative conditions of illumination for the two, the gray may reflect much more white light from its surface than the white.

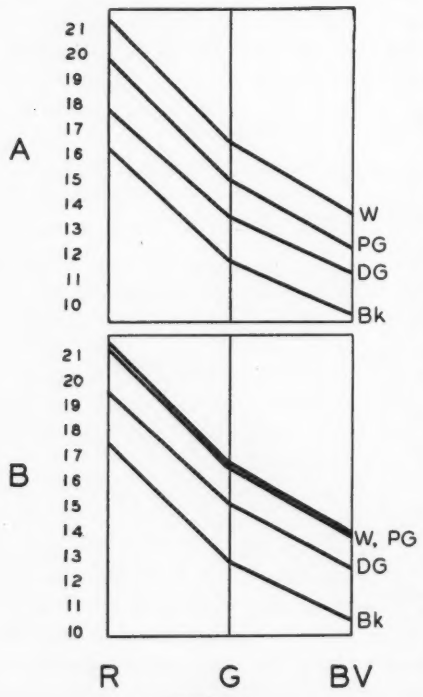

FIGURE 1

Tint-photometer readings (ordinates) with the three color-filters (R, G, BV), for melanin samples from fishes kept in the white, pale gray, dark gray and black aquaria in the "A" (brightly lighted) and "B" (dimly lighted) cabinets.

It may be well to remind the reader at this point that the adjustment of fishes and amphibia to their backgrounds depends upon two processes: (1) a relatively rapid one, resulting from the concentration or dispersion of pigment particles already contained in the chromatophores; and (2) a slower process, involving changes in the actual amount of pigment present. It is with the second class of phenomena that we are concerned here.

We wish to acknowledge important assistance received in the course of these studies, notably from Professors R. A. Gortner, R. S. Woodworth, Eric Ogden, D. L. Fox, G. E. K. Branch and J. W. Ellis.

The fish used in the present series of experiments was a marine species, the long-jawed goby, *Gillicthys mirabilis*. Specimens of this fish (mean weight 28.2 grams) were placed in eight experimental aquaria (39 x 25 x 25 cm.), 12 fishes in each aquarium. Four of these aquaria were kept in each of two light-proof cabinets, one of which ("A") was lighted by two 200-Watt incandescent lamps, the other ("B") by two 10-Watt lamps. These cabinets were lighted continuously, day and night. Fishes were subjected to these conditions from December 12, 1935, to March 9, 1936. They were fed three times a week upon polychaete worms. The water temperature during this period ranged from 13° to 16°C.

The relative light intensities just above the aquaria, measured with a Weston foot-candle meter, were approximately 140 and 4 foot-candles, respectively, for the two cabinets. Since, however, the aquaria were covered with heavy wire screen to prevent the escape of the fishes, the amount of light actually received by the animals was considerably reduced. Thus, when the meter was placed at the bottom of the (empty) white aquaria and the screens placed in position, the readings were about 65 and 1.5 foot-candles, respectively, for the "A" and "B" cabinets. Accordingly, the fishes in the former cabinet received roughly 40 times as much light from overhead as those in the latter.

The glass walls of the four aquaria in each cabinet were painted white, pale gray, dark gray and black, respectively. The albedos of the surfaces, determined photometrically,[5] were approximately 36, 15.9, 5.9 and 0. That is, the inner surfaces of the white aquaria reflected about 36 per cent of the light received by them, the pale gray 15.9 per cent, etc. Regarding the black aquaria, it must be stated that, when in actual use, their albedo was somewhat higher than zero, owing to light-colored detritus which accumulated upon the bottom, despite rather frequent removal. And it must be further added that the fishes themselves, which were never completely black, constituted a significant part of one another's background. These same circumstances doubtless affected the albedo of the backgrounds in all the aquaria.

After 87 days' subjection to the conditions described, the fishes were killed, weighed and put through a course of treatment, the object of which was to get rid of as much of the body substance as possible, without losing any significant fraction of the melanin, and without producing any colored substances which might augment the apparent quantity of melanin at the end.

To summarize the chief steps in this process, which was the outcome of some months of experimentation, (1) the fishes were ground in a ball-mill; (2) the resulting material was subjected to Soxhlet fat-extraction; (3) it was decalcified in dilute HCL; (4) it was boiled in 0.2 per cent NaOH; (5) the resulting colloidal solution was neutralized with HCL and (6) the

precipitate collected by centrifuging; (7) the latter material was subjected to digestion in a pepsin-HCL mixture for 48 hours; (8) recovered by centrifuging and (9) redissolved in dilute NaOH. This solution contained all or nearly all of the melanin, plus a certain proportion of colorless organic substances (protein derivatives?) which it was found impossible to remove without introducing even greater errors.

It proved impracticable to make direct colorimetric readings of the brown liquid thus obtained, owing to an unavoidable slight cloudiness, which naturally affected the transmission of light. Nor were we, for various

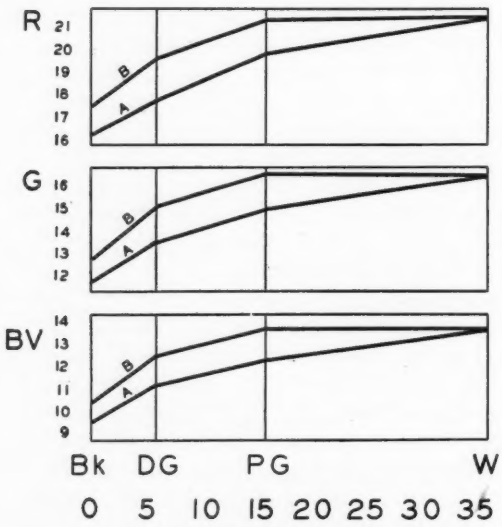

FIGURE 2

Based upon same tint-photometer readings as figure 1, but data differently arranged. Abscissa scale indicates albedos of the various tanks.

reasons, able to make reliable gravimetric determinations of the melanin, although much time was devoted to an endeavor to do so.

The procedure finally adopted[6] involved the addition of a pure white substance (magnesium carbonate) to the liquid ("9" above), drying and grinding this, mixing samples of it with cedar-wood oil, transferring the resulting paste to special cells and taking photometric readings of it as in the case of any opaque substance.[7]

It should be stated that all of the chemical reagents and other substances employed throughout this procedure were used in amounts proportional to the original weight of the fishes.

Readings obtained in this way denote percentages of light reflected from the given sample in comparison with that reflected from a block of dry magnesium carbonate, used as a standard. Three color filters (red, green and blue-violet) were used successively and readings recorded from each. Figure 1 is based upon these readings, for the fishes from the four aquaria in the "A" and "B" cabinets, respectively.

In figure 2, the same data are presented in a different form. Here it is possible to compare the readings for the "A" and "B" series, as well as to obtain a better picture of the relations between albedo and pigmentation. It should be borne in mind that high readings denote low melanin values and vice versa.

Such readings are not, of course, convertible into actual quantities of melanin. But since the proportions of melanin in these samples of the mixture with magnesia correspond in every case to those present in the various lots of fishes from which they were taken, their relative values are presumably correct. In order to determine the relative melanin contents of these different lots of fishes, samples of the magnesia preparation from the darkest fishes (those kept on black) were further diluted with magnesia and readings taken of the various resulting percentages of "black" material from 60 to 100. Thus, "white" fishes from the "A" cabinet were found to give photometric readings corresponding to a mixture containing about 73 per cent of material from the "black" fishes.

Table 1 presents both the mean tint-photometer readings for the samples of differing history, and the computed values of melanin, regarded as percentages of the amount present in the "A" black series.[8] In figure 3, these latter values are plotted in graphic form against the albedos of the respective tanks.

As already noted, the actual quantity of melanin is not indicated by any of the foregoing figures. We have, however, made comparisons of our material with various similar mixtures of magnesia with melanin of known purity, which was kindly furnished us by Professor R. A. Gortner. If we may legitimately compare the melanin from our gobies with that derived from the feathers of black Minorca fowls, we find that our "black" gobies contained about 43 mg. of melanin per 100 grams weight of the fishes. While it is not yet certain that we are dealing with identical substances in these two widely different animals, we have some further evidence that the quantity of melanin contained in these fishes is of this general order of magnitude.

The values given in the preceding table and plotted as curves (figures 1, 2 and 3) represent the means of three samples of three fishes each, into which each lot of fishes was subdivided. (A fourth lot, "no. 1," had been sacrificed in experimenting upon technique.) These samples, comprising fishes of supposedly identical history, none the less exhibit considerable

differences among themselves, as is shown by the circles in figure 3. To what extent these differences are due to actual individual differences among the fishes, and to what extent they are due to unperceived minor variations in procedure, we cannot state. Whatever their cause, they are not sufficient to obscure certain general trends, which are revealed both when the fishes from the two cabinets are compared with one another, and when the fishes from the various aquaria in each cabinet are compared with one another. This may be stated with confidence, without our resorting to a mathematical calculation of probabilities which, in the present case, would be very difficult.

Several facts revealed by these studies deserve particular comment:

(1) We find a rather regular gradation in melanin content among the fishes in the four aquaria in the "A" cabinet, and, except for one striking

TABLE I

		TINT PHOTOMETER A	TINT PHOTOMETER B	% "BLACK" A	% "BLACK" B
R	W	21.559	21.622	72.5	72
	P.G.	20.016	21.487	79.5	73
	D.G.	17.895	19.767	92.5	81
	Bk	16.439	17.688	100.5	94
G	W	16.622	16.731	73	72
	P.G.	15.141	16.717	80.5	72
	D.G.	13.649	15.313	89.5	79.5
	Bk	11.917	12.921	100.5	94
B-V	W	13.787	13.893	74.5	74
	P.G.	12.463	13.938	82	73.5
	D.G.	11.415	12.661	90	81
	Bk	9.816	10.629	101	96
Means	W			73.3	72.7
	P.G.			80.7	72.8
	D.G.			90.7	80.5
	Bk			100.7	94.7

exception among the "white" samples, a similar gradation in the "B" cabinet.

(2) In the "A" cabinet, at least, the melanin content proved to be nearly inversely proportional to the logarithm of the albedo of the background. Plotted on semi-logarithmic paper, the values for "white," "pale gray" and "dark gray" fall into an almost perfectly straight line. That for "black" would fall into the same line if the albedo of this tank were to be regarded as 2.3 instead of 0 (see p. 213). While we do not wish to generalize at this stage of our work, it is interesting to point out the

approximation of this result to Fechner's formulation of Weber's law, viz., $S = C \log \Sigma$, in which S = the magnitude of the sensation and Σ the magnitude of the stimulus.

(3) There are also fairly consistent, though smaller, differences between the values for the "A" and "B" cabinets. This difference is more than 7 per cent, the "A" fishes showing higher values than the "B" ones. It will be recalled that the relative illumination of these two cabinets was in the ratio of about 40 to 1. As compared with this, the mean difference between the lots from the successive aquaria in the "A" cabinet was more

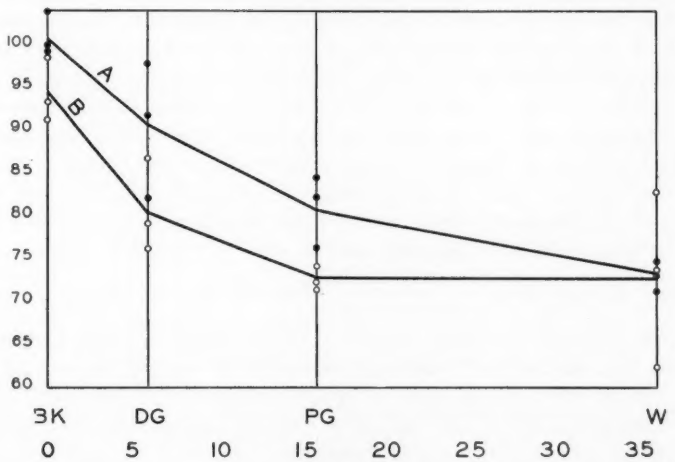

FIGURE 3

Curves showing the mean relative melanin values of the four lots in each cabinet, taken as percentages of the value for "A black" fishes, regarded as 100. Ordinates indicate the percentages, abscissas the albedos. The circles show the values for the three samples of three fishes each, into which each of the lots in any given aquarium were divided. Solid circles represent those from the "A" cabinet, open circles those from the "B" cabinet.

than 10 per cent, while the albedos of the white and pale gray aquaria stood in the ratio of 6.3 to 1, and those of the pale gray and dark gray stood in the ratio of 2.7 to 1.[9] Thus, the effects of differences of incident light, within the limits here employed, were small in comparison with the effects of differences in albedo.

(4) Both the differences observable in the melanin content of the various lots of fishes in each cabinet and those between the fishes of the two cabinets as a whole are in the same direction as the visible differences in shade which had already been noted in the living animals at the close of the experi-

ment, these being placed together on a white background for comparison. The differences in melanin were, however, small as compared with the very pronounced differences in shade shown by the living fishes.[10] It is evident that the greater part of these visible differences was due to differences in the state of aggregation or dispersion of the melanin within the chromatophores.

While our experiments were under way, experiments in some respects similar to ours were reported by F. A. Brown[11] at the University of Illinois. Brown's experimental set-up resembled ours in essentials, but his objective was somewhat different, since he was concerned not with the quantity of melanin but with the surface extension of the pigment in the chromatophores. Furthermore, Brown's fishes were kept for relatively very brief periods (two to three hours) under the various conditions of illumination. He reports (1) that at certain very low light intensities (0.000005 to 0.25 foot-candle) in fishes kept upon identical backgrounds, "the average diameter of pigment masses is directly proportional to the logarithm of the light intensity," and (2) at a certain point (1.75 foot-candles) "the fish becomes maximally dark upon a black background, and the average diameter of melanin masses appears to vary in a directly proportional fashion with the ratio, $\frac{\text{incident light}}{\text{reflected light}}$, which reaches the eyes of the fish."

We have no data for melanin values at the feeble illuminations considered in the first part of the foregoing statement. As regards the second, we have shown that the quantity of melanin produced (or retained) varied more nearly according to a logarithmic relation than a linear one. There is, however, no necessary contradiction between Brown's results and ours.

We are aware of no previous attempt to measure the melanin of fishes and only one other attempt to evaluate the differences of melanin content in any animal capable of adjusting its shade to that of its background. Vilter has published a brief account[12] of the results of extracting the melanin from two axolotls, which had been kept for 17 months on black and white backgrounds, respectively. He states that the former animal yielded four times as much melanin as the latter. Owing to the small number of individuals employed, and in the absence of any statement regarding the intensity of illumination, or of the method employed in measuring the melanin, Vilter's figures can hardly be regarded as precisely quantitative.

[1] Sumner and Keys, *Physiol. Zoöl.*, **2**, 495–504 (1929). Some highly interesting experiments bearing upon the same general problem are discussed by Burkamp, *Zeitschr. Psychol. Physiol. Sinnesorgane*, II Abt., 55 Bd., S. 133–170 (1923).

[2] That the stimuli responsible for these changes are visual ones is now too well established to demand discussion here.

[3] Keeble and Gamble, *Phil. Trans. Roy. Soc.* (B), **196**, 295–388 (1904); Sumner, *Jour. Exp. Zoöl.*, **10**, 409–505 (1911).

[4] This resolves itself, of course, into the relative stimulation of different parts of the retina (Sumner, *Biol. Bull.*, **65**, 266–282 (1933)).

[5] The Ives tint-photometer and Weston foot-candle meter were used for these determinations, which cannot be regarded as more than rather rough approximations.

[6] We owe to Professor Eric Ogden, of the University of California, the suggestion which led us to adopt this procedure.

[7] The Ives tint-photometer was used for this purpose. (See Sumner, *Jour. Mammalogy*, **8**, 177–206 (1927); Sumner and Fox, *Jour. Exp. Zoöl.*, **66**, 263–301 (1933)).

[8] The presence of values greater than 100 is due to the fact that the "black" figures in the table are based upon samples which were not identical with those which were used as our standard.

[9] The corresponding ratio of dark gray and black would be impossible to state, since the albedo of the latter is not exactly known (cf. p. 213). If a value of 1 be assumed, the ratio would be 5.9 to 1.

[10] Small also in comparison with the effects earlier observed in *Lebistes*. (Sumner and Wells, *Jour. Exp. Zoöl.*, **64**, 377–403 (1933).

[11] *Biol. Bull.*, **70**, 8–15 (1936).

[12] *C. R. Soc. biol.*, **108**, 774 (1931).

A REMARKABLE PARALLELISM

BY RICHARD GOLDSCHMIDT

DEPARTMENT OF ZOÖLOGY, UNIVERSITY OF CALIFORNIA

Communicated March 12, 1937

In a series of papers, partly in press,[1] I have reported upon investigations of development and genetics of the vestigial mutant in *Drosophila*. The decisive facts are: The wing of this mutant and its series of multiple alleles is a clipped wing. It looks as if parts from a whole wing were clipped after the whole wing was formed. This clipping begins at the tip and proceeds in a perfectly regular fashion in the ascending series of multiple alleles until, in the highest member No-wing, hardly any stumps are left. This aspect of the series already forces one to conclude that the spatial determination of the wing, the position of veins, etc., was determined before the process controlled by the *vg*-genes set in, which results in the clipping. The different conditions of the wings, therefore, cannot be described as growth stages, as has been done in many quantitative investigations. The study of the development revealed that this *a priori* consideration was correct. We divide for descriptive purposes the grades of scalloping, beginning with a small nick and ending in the stumps of No-wing, into eleven classes, of which *X* represents the type vestigial and V–VI about notched toward snipped. It was shown that in the classes I–V/VI the wing is still perfectly normal at the time of pupation. Afterward a degeneration or lysis of wing epithelium begins which destroys progressive amounts of tissue from the

tip on. This process stops when the moment of histological differentiation has come, and the rest of the wing develops normally. The destroyed parts of wing area are the scallopings. The time of onset of the destructive process is now exactly correlated to the amount of scalloping produced, i.e., the higher alleles (or corresponding heterozygotes and compounds) produce an earlier and earlier onset of the degenerative process. With a phenotype between classes V and VI the visible onset is still after pupation. In the higher classes it occurs in the imaginal discs (as extrapolated from the facts of later development by the author and actually found by Auerbach[2]) and in the highest, *vg* and *Nw*, in the very young imaginal disc. But, also, in these higher grades the destruction is not yet finished at the time of pupation but proceeds considerably up to the last possible moment.

It has been known for a long time that temperatures at the limit of the vital range (29–31°C) make a genetical vestigial wing more normal-like and that this effect is strictly proportional to the increase in temperature within the above-mentioned range, and further that the phenotypes thus produced run through the whole series of the phenotypes of the *vg*-alleles. The author could then produce a phenotypic effect representing classes I–VI of scalloping from wild-type flies by the action of temperature shocks and the same has since been performed by Friesen[3] employing x-rays. The author could furthermore isolate a number of dominance modifying genes (dominigenes) which shift the otherwise wild-type *vg*/+ heterozygote toward the recessive type, producing grades I–IV of scalloping in an additive way. It was then shown that the same dominigenes act also upon the homozygous alleles and the different compounds, shifting their phenotypes in a perfectly orderly series toward the higher classes. All these effects take place in the same way, namely, changing the phenotype in a linear order through the different phenotypes I–XI, identical with the known phenotypes of the allelic series. The underlying physiological system must therefore be a simple one which allows shifts only in a plus or minus direction.

These facts may be described in terms of gene action in two ways: Either the genes control the production of some substance which leads to wing degeneration when a certain threshold of concentration is reached and the time at which it is reached is a function of the different conditions of the gene, called alleles. Or the genes control the production of a substance which is necessary for normal growth (a growth-promoting substance) and which has to be present in a minimum concentration to enable development of the wild-type wing. The different alleles produce an insufficiency of this substance which, when fallen below a certain threshold, will lead to an insufficiency disease, degeneration. Originally I preferred the first interpretation, but later I adopted what seemed to be a more plausible description in terms of insufficiency of a growth substance.

In both cases the facts and their interpretation might be represented by a graph (Fig. 1). The normal curve on top represents the curve of growth of the normal wing, and in order not to draw an extra curve, may also be taken as the curve of production of the growth-promoting substance, using the second interpretation. The abscissa indicates the instars of development, including pupation and at the time Det final determination occurs. The lines parallel to the abscissa indicate the minimum threshold below which there would be insufficiency of this substance. In the different alleles destruction = insufficiency sets in at different times and destroys wing tissue. The size of the wing at different times, as actually found, may then be plotted upon the ordinates for the different alleles, and the time

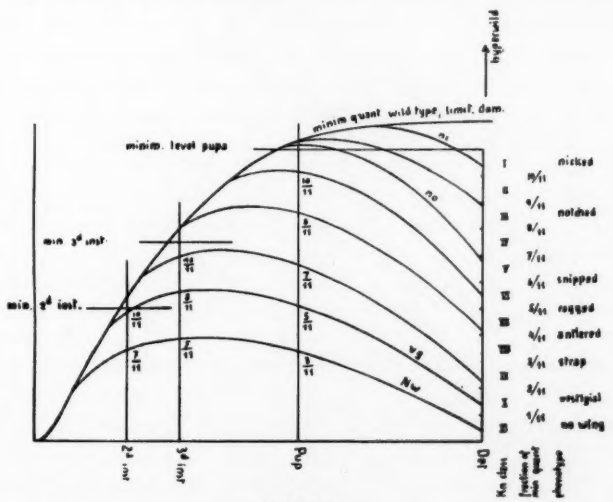

FIGURE 1

of onset of degeneration indicates when the substance falls below the threshold. Thus the curves for first normal growth and subsequent increasing degeneration of wing area are found as represented in a rough approximation in the graph. Within such a system, of course, whatever agent affects the decisive process (temperature, modifiers) will have either a plus or minus effect, i.e., will shift the curves from one level to the other.

I have since become acquainted with the work of Krueger[4] on the kinetics of the bacteriophage and was struck by the tremendous similarity of both analyses. Krueger found the following: If a mixture of phage and growing bacteria is kept, the increase of phage depends upon the bacterial growth in so far as it does not occur without the latter. But the rate of increase of phage is considerably greater than the rate of bacterial repro-

duction and therefore the ratio of phage to bacteria is constantly increasing. When a certain ratio in favor of phage is reached—the lytic threshold—lytic destruction of bacteria begins and proceeds rapidly to completion. *The time of onset of this destruction is proportional to the initial concentration of phage in the mixture.* Phage may be inactivated and reactivated by different ions. In addition, in the presence of manganese the lytic action begins earlier; this is solely an effect upon the threshold, i.e., the quantity of phage/bacterium requisite for lysis. Figure 2 represents Krueger's graph for these experiments, which may be said to be also a perfect model

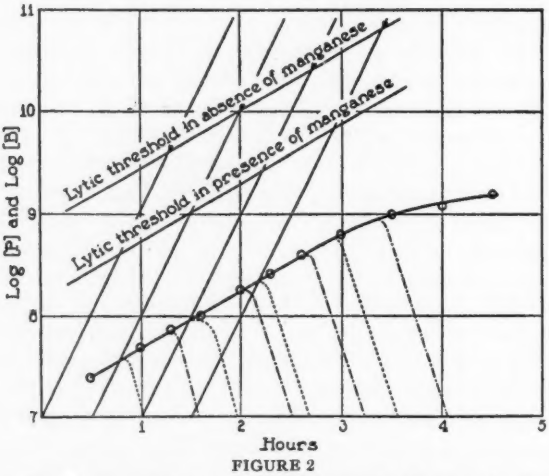

FIGURE 2

Graphic representation of bacterial growth, phage production and lytic ratios with and without $MnCl_2$. Originating in the bacterial growth curve (○) are the curves of bacterial lysis(. . . .) with $MnCl_2$ and adjacent broken lines the identical mixtures without $MnCl_2$ (after Krueger).

of the vestigial case. The growth curve of bacteria corresponds to the normal wing growth in *Drosophila*. The same curve might also represent the increase in phage if it were proportional to the bacterial growth, as it has been tacitly assumed in the vestigial case for the production of the decisive stuff. This assumption, which works in case of a growth substance, would not work in case of production of a lytic substance, as we must account for the different onset of action. In Krueger's diagram the different times of onset of lytic action and the consequent dropping of the curves for bacterial growth are visible and their similarity to the curves for wing growth is obvious. Here we find, in addition, the steep curves for growth of the phage drawn for four different initial concentrations. Their intercept

with a line parallel to the growth curve, representing the assumed ratio of quantity phage/bacterium necessary for lytic action, is then the lytic threshold. In the vestigial curve this threshold has been indicated directly, but it might as well be represented as in Krueger's graph. The upper parallel indicates the lytic threshold in the presence of Mn. A parallel to this would be the action of the dominigenes (or temperatures) upon the *vg*-curves.

It is clear that we could use the graph for the phage also to describe the vestigial case, namely, under the former assumption that the genes produce here a lytic substance which acts upon the wing as soon as a threshold value is reached. The main curves of Krueger's graph would then represent the growth and subsequent lysis of the wing; his curves for phage production would be the curves for production of the lytic substance; the initial concentration of phage would be the different alleles of *vg*, and the manganese curves would represent the action of modifiers.

The remarkable fact about this parallel is that here the kinetics of a case have been studied which in every detail fit the kinetics of gene action as derived from a special case (and, incidentally, is identical with the conclusions derived from the author's former work on gene action). I should, however, not go further and draw parallels between genes and phage, as has occasionally been done. The parallel seems to be perfect, so far as kinetics are concerned, but it does not contain any necessary information about the nature of the thing which acts according to initial concentration.

The present idea regarding the nature of the phage seems to be that it is a protein (produced by the bacteria) with catalytic properties. Such properties are also supposed to be typical for the gene, whatever it is. But here the comparison ends for the present.

[1] "Gen und Ausseneigenschaft. I, II." *Zeitschr. ind. Abstl.*, **69** (1935); III. *Biol. Centrbl.* (1935); IV–VI. *Univ. Calif. Publ. Zool.* (1937).

[2] *Trans. Roy. Soc., Edinburgh*, **58** (1936).

[3] *Roux' Arch.*, **134** (1936).

[4] *Phys. Rev.*, **16** (1936) (here complete list of literature).

THE RELATION OF GENE TO CHARACTER IN QUANTITATIVE INHERITANCE

BY EDMUND W. SINNOTT

DEPARTMENT OF BOTANY, BARNARD COLLEGE, COLUMBIA UNIVERSITY

Communicated February 18, 1937

The multiple-factor theory of the inheritance of quantitative traits postulates the operation of many genes in determining a given character. It seems usually to have been assumed that each gene produces a constant and specific effect on the phenotype and that these effects are arithmetically cumulative, so that a segregating F_2 population should show a symmetrical distribution. East[1] long ago discussed the probability that gene action is accelerative and that it varies depending on the rest of the genotype. The same idea has occasionally been expressed by other geneticists. Indeed, everything known as to the nature of gene action is quite opposed to the naive idea that a given gene will always contribute a constant absolute amount to the phenotype. No alternative hypothesis seems to have been formally set up, however, until Rasmusson[2] presented evidence that multiple genes interact in such a way that their individual expression may be greatly modified, "the visible effect of a certain factor being smaller the greater the number of factors acting in the same direction." Something like the "law of diminishing returns" is thus assumed to be operative. Genes affecting quantitative traits should thus be geometrically rather than arithmetically cumulative.

Evidence obtained by the writer as to the inheritance of fruit weight in *Cucurbita Pepo* supports the hypothesis of the geometric action of genes affecting size, but indicates that each progressive genic increment has a progressively *greater* rather than smaller effect. The character of the genic effect can best be determined from a study of the distribution of an F_2 population as to a given quantitative trait. If this is symmetrical about its mean, one must conclude that each gene makes a constant absolute contribution to the phenotype and that these are arithmetically cumulative. If the distribution is consistently asymmetric or skewed, however, this may be considered as evidence that the genes are interacting and producing a geometric or multiplicative effect. The type of interaction suggested by Rasmusson would result in a curve skewed negatively, since genically equivalent classes would be separated by progressively smaller increments. If the hypothesis advanced in the present paper is correct, however, precisely the opposite result should follow and the curve be skewed positively, since equivalent classes would be separated by progressively larger increments.

The distribution of fruit weight was studied in 43 F_2 populations, including over 4500 individuals, from crosses between pure lines differing in fruit weight. In 41 of these there is definite positive skewness, the mean being larger than the median. A typical example is shown in figure 1A, consisting of 244 individuals and with a skewness of +0.487. Distributions similar to these may be found in many segregating populations of plants and animals described in the literature of quantitative inheritance.

These facts can well be explained if it is assumed that the effect of a given gene increases in proportion to the effectiveness of the rest of the genotype. A gene may thus be thought of as adding a definite percentage to what the rest of the genotype produces, rather than a definite absolute value. A series of segregating genes of this sort (doubtless differing in

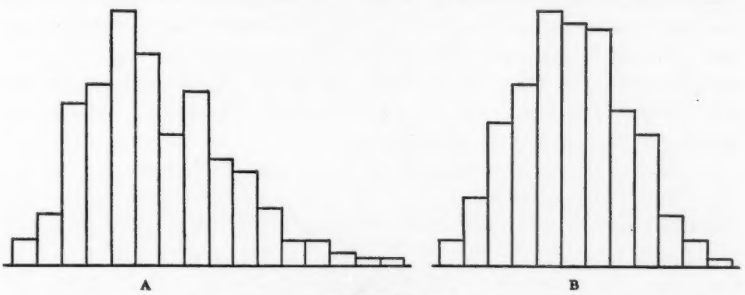

FIGURE 1

A, segregating population of 244 individuals plotted in arithmetically equal classes; skewness +0.487. B, the same population plotted in equal logarithmic classes; skewness −0.057.

their individual effectiveness) would produce an asymmetrical, positively skewed population.

If this conception of gene action is correct, the division of a population into arithmetically equal classes, as is usually done, evidently gives an erroneous idea of the segregation which is taking place and should be abandoned for a method in which a given gene has the same quantitative representation wherever it occurs. The most obvious suggestion for accomplishing this is to make the classes logarithmically equal, by plotting the logarithms of the values rather than the values themselves. The F_2 populations described above, segregating for fruit weight, were thus plotted, and in almost every case a very much closer approach to symmetrical distribution resulted. This treatment of the same population shown in figure 1A is represented in figure 1B, the skewness here being only −0.057. The value of such a method of analyzing the inheritance of quantitative characters is obvious.

In the majority of populations thus logarithmically plotted, some negative skewness results, which occasionally becomes considerable. This seems to be related to the curious fact that in the pure lines and the F_1 progenies studied, where all the variation is presumably environmental, the distribution in most cases is essentially symmetrical. Skewness seems to be associated primarily with the variability resulting from genetic segregation. Logarithmic plotting of a population symmetrically distributed in arithmetic classes will evidently cause it to skew negatively, and the presence of negative skewness in some F_2 populations under logarithmic plotting may therefore be an indication of the existence there of a certain amount of environmentally produced variation.

The geometrically progressive character of genic effects may be explained by the general facts that growth is exponential and that genes control the rates of growth processes. It might naturally be expected that successive increments of growth, equal in time or as to their genetic bases, would be progressively greater in absolute value. On this hypothesis, however, there would seem to be no reason why genetically pure material, subject only to environmental variability, should not show a similar asymmetric distribution.

Another explanation is suggested by a histological study of fruit development recently made in this material. Fruit size differences at maturity are due to several distinct and apparently independent processes. The production of new cells by division, the expansion of cells to their final size, the attainment of reproductive maturity (which affects the duration of cell division), and the development of the secondary cell wall (which affects the extent of cell enlargement) are four such processes. More detailed developmental analysis would doubtless discover many others, with varied genetic bases. It is thus obvious why most traits of size in such material should be determined by many genes. If distinct genes control the rates of distinct processes, and if the processes are independent developmentally (and there is evidence that both these assumptions are true, at least in part), it is clear that the interaction of one process upon another will be multiplicative. If one gene tends to double the number of cells, for example, and another to triple their volume, the combined effect of the two will be to multiply the volume of the fruit by six times. A series of such diverse genes would thus produce through their interaction a geometric series of effects and an F_2 population showing the positive skewness which has here been reported.

Independent genes affecting the *same* process, when brought together, may very possibly show a progressive reduction in their individual effect, as Rasmusson suggests; but such cases are either rather rare or are masked by the other and commoner type of interaction. Allelomorphic genes, especially, might show diminishing effects.

The apparent rarity of geometric effects in environmental variability is perhaps due to the fact that some environmental factors are stimulatory and some inhibitive in effect. A geometrically progressive inhibiting factor would tend to result in negative skewness. Possibly the apparent absence of skewness may be the result of the cancellation of one type of effect by the other. Such speculations are rather fruitless in the absence of more definite evidence.

Summary.—Evidence is presented that in the inheritance of fruit size in *Cucurbita Pepo* the genes have a geometrically cumulative effect, resulting in a positively skewed distribution in F_2. Much more nearly symmetrical distributions are obtained by plotting in logarithmically equal classes. The suggestion is made that this type of gene action results from the fact that many processes, independent genetically and developmentally, are concerned in the determination of size, and that these interact in a multiplicative fashion.

[1] East, E. M., *Bot. Gaz.*, **55**, 177–188 (1913).
[2] Rasmusson, J., *Hereditas*, **18**, 244–261 (1933).

THE INSTANTANEOUS VISUAL THRESHOLD AFTER LIGHT ADAPTATION

BY SELIG HECHT

LABORATORY OF BIOPHYSICS, COLUMBIA UNIVERSITY

Communicated March 11, 1937

One of the satisfactions in applying a theory comes from its unexpected illumination of data which, though well known, have long remained obscure in their interpretation and unrelated to the rest of the field. The measurements of visual intensity discrimination are a case in point. Though accumulated[1,2,3] over a period of 75 years, they have only recently been recognized[4] as falling into (*a*) the physiological scheme of the duplicity theory[5] which separates vision into cone function at high illuminations, and rod vision at low illuminations, and (*b*) the mathematical pattern which describes the first steps of photoreception in terms of photochemical and dark reactions arranged as a cyclic pseudoreversible system.[4] The familiar data of Blanchard[3] on instantaneous thresholds bear on precisely the same ideas, but have failed of interpretation until now.

In Blanchard's measurements the eye is adapted to a given intensity by looking for some time at a large suitably illuminated area. At a specific moment this adapting light is cut off, and the minimum illumination is

determined at which a centrally fixated 5° area becomes visible at the same moment. This requires several trials, each preceded by proper adaptation, until the precise intensity is found which renders the central area just visible instantaneously on shutting off the adapting light.

Blanchard made these measurements over a large range of adapting intensities, using white light as well as the red, yellow, green and blue portions of the spectrum isolated by roughly monochromatic filters. He found that the instantaneous threshold intensity increases with the adapt-

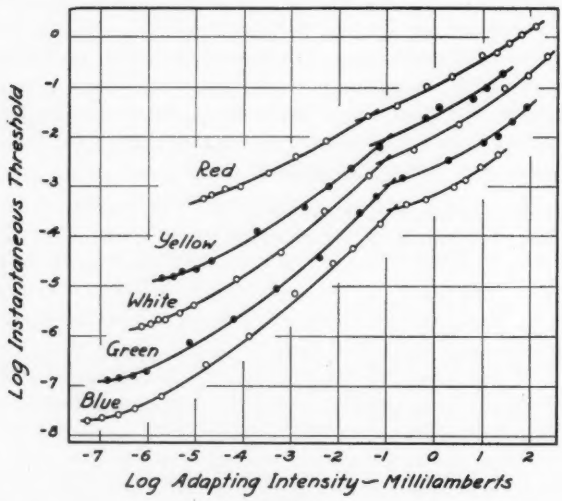

FIGURE 1

Blanchard's measurements of the instantaneous threshold following light adaptation to various intensities and colors. The adapting intensity axis is the same for all the data. The numbers on the instantaneous threshold axis apply only to the white data because, to avoid confusion in plotting, the yellow data have been displaced upward 0.5 log unit and the red 1 log unit; the green data have been displaced downward 0.5 log unit and the blue 1 log unit.

ing intensity, and states that on a double logarithmic plot the data, except at the extreme intensities, lie on practically straight lines which he represents by the appropriate equation for a straight line. Blanchard notes that the data for the different colors are very similar at moderate and high brightnesses, but that they diverge at the lowest brightness, and he correctly attributes the latter fact to the Purkinje phenomenon.

Blanchard adds the pertinent observation that "all the color curves except red show a decided dip, indicating a depression of sensibility, in a

region roughly between 0.01 and 1 millilambert, a region corresponding to about the average range of interior brightnesses at night," but fails completely to appreciate the meaning of his observation. I have replotted his data in figure 1, from which the significance of his "dip" is at once apparent in terms of the duplicity theory. The reason that Blanchard speaks of the region of the rod-cone transition as a dip rather than as a rise is that he actually plotted the negative logarithm of the instantaneous threshold (calling it the sensibility), and thus his curve is reversed in direction of the ordinates.

Figure 1 shows that with red light the measurements form an almost continuous function. Indeed, when plotted by themselves, the red data may easily be fitted with one line; it is only when the data for the other colors are added that the double nature of the function becomes apparent because the inflexion point in the red data, though gentle, corresponds so obviously with the point of striking change in slope shown by the yellow, white, green and blue data. Judged by the transmission data published by Blanchard, his red light begins at 600 mμ; it would therefore stimulate the rods as well as the cones in this 5° centrally fixated area.

In figure 1 it is apparent that the high intensity portions for all the colors and for white are much the same, indicating their common origin in cone function. Except for red, the low intensity sections drop sharply away, also in a form which is approximately the same for all the measurements. Note particularly that the low intensity rod section is largest for blue light and becomes smaller through the spectrum to the red, as is to be expected from the spectral sensibility curves of the rods and cones as separately determined. Attention has already been called to this situation in dark adaptation,[6] in intensity discrimination[4] and in flicker,[7] and it is unnecessary to labor the point here. The instantaneous threshold measurements fit directly into the pattern of the duplicity theory, and serve to strengthen the force and validity of that generalization.

These data of Blanchard's are significant beyond their rod-cone interpretation because of their quantitative numerical relations. In this respect, they are like the data of intensity discrimination and flicker, since they fall into the photochemical scheme proposed for the initial events in the photoreceptor process.[4]

In general terms this scheme supposes that the photoreceptor process is composed of (*a*) a primary photochemical reaction in which an inactive photosensitive material absorbs light and is converted by it into one or more active products; (*b*) a primary dark reaction which maintains the supply of sensitive material, since otherwise it would be used up and vision would cease; and (*c*) a secondary dark reaction in the course of which the active photoproducts of the primary photochemical reaction undergo some

change, the end result of which is an impulse from the receptor cell to the attached nerve.

Of course the photoreceptor system is more complicated than this, and to describe some visual data it may be necessary to make further suppositions. For dealing with the present data, however, it is not necessary to consider even the whole of the suggested system; we need to study only the primary light and the primary dark reactions. In such a simplified arrangement, the sensitive material S is changed by light into the photoproducts $P, A, \ldots$ some of which may under proper conditions reunite to form the sensitive material from which they were derived.

Let the total initial concentration of S be a, and the concentration of $P, A, \ldots$ at the moment t be x. Under the influence of light of intensity I the photochemical process will have a velocity

$$(dx/dt)_1 = k_1 I\,(a - x)^m \tag{1}$$

where k_1 is the velocity constant which includes the absorption coefficient, and m is the apparent order of the reaction. The simultaneous dark or thermal reaction which furnishes the sensitive material S from the main photoproducts $P, A, \ldots$ has a velocity

$$(dx/dt)_2 = k_2 x^n \tag{2}$$

where k_2 is a velocity constant and n is the order of the reaction. If the light I continues to illuminate the eye, the two velocities become equal. Putting equation (1) equal to (2) gives

$$KI = x^n/(a - x)^m, \tag{3}$$

the familar stationary state equation in which $K = k_1/k_2$.

After producing this stationary state, the light I is sharply cut off; the instantaneous threshold I_0 is then the least intensity of light which is visible to the eye at the very first instant of darkness. At this first moment, before the stationary state has undergone any change, the receptor system becomes exposed to the instantaneous threshold intensity I_0. The velocity of the resulting photochemical action is

$$(dx/dt)_{I_0} = k_1 I_0\,(a - x)^m \tag{4}$$

where $(a - x)$ is still the concentration of sensitive material in the stationary state. In treating intensity discrimination[4] we assumed that to produce a minimal visual effect, the initial velocity of this photochemical reaction is constant and equal to c'. To make the same assumption here means putting equation (4) equal to c', and gives

$$I_0 = c'/k_1\,(a - x)^m. \tag{5}$$

For this expression to be useful with measurements, we need to substitute

for $(a - x)$ the value of the corresponding adapting intensity I in the stationary state equation (3). When $m = n = 1$, equation (3) yields $(a - x) = a/(1 + KI)$, which when introduced in (5) gives

$$I_0 = c(1 + KI) \tag{6}$$

where $c = c'/k_1a$. When $m = n = 2$, equation (3) yields $(a - x) = a/[1 + (KI)^{1/2}]$; after this value is put into equation (5) we get

$$I_0 = c[1 + (KI)^{1/2}]^2 \tag{7}$$

where $c = c'/k_1a^2$.

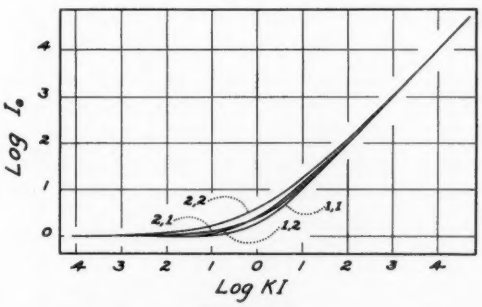

FIGURE 2

Theoretical relation between the instantaneous threshold I_0 and the adapting intensity I in terms of equations (6) and (7). For completeness the curves corresponding to $m = 2$, $n = 1$, and $m = 1$, $n = 2$, have been added. Because they are plotted on a double logarithmic grid, the shape of the curves is invariant in relation to the constants c and K which merely determine the position of the curves on the ordinates and abscissas, respectively.

The curves which correspond to equations (6) and (7) are plotted in figure 2; the two curves showing the situation when $m = 1$, $n = 2$, and $m = 2$, $n = 1$ are also included. The curves are plotted as log I_0 against log KI, a procedure which renders their form invariant. The shape of the relation between the two variables remains the same regardless of the constants involved; changes in the constant c merely move the curve as a whole along the ordinates, while changes in K move the curve along the abscissas.

Examination of the theoretical curves in figure 2 shows their similarity to Blanchard's actual measurements in figure 1. However, a direct comparison of theory and data is not immediately possible because the measurements were made with the natural pupil, which varies enough over this

large range of intensities to distort the actual relationship. Fortunately Blanchard records his own pupil diameter at different illuminations. The corresponding pupil areas can then be corrected for their relative efficiency in brightness transmission as determined by Stiles and Crawford,[8] and the resulting effective pupil areas used for converting millilamberts, which are brightness units, into effective photons which are retinal units.[9] Since the instantaneous threshold represents a measurement made at the first instant after light adaptation when no pupil change can as yet have occurred, the same effective pupil area correction applied to the adapting intensity must also be applied to the corresponding instantaneous threshold intensity.

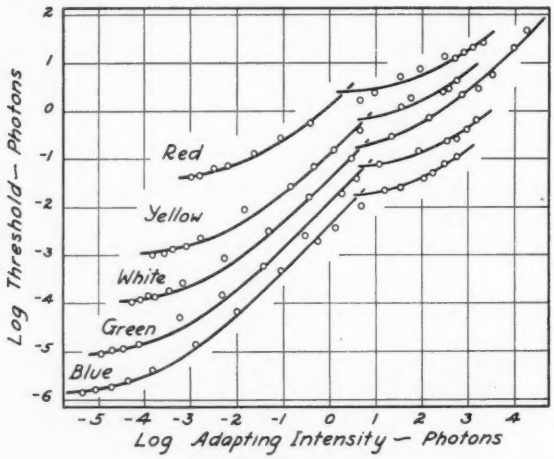

FIGURE 3

Blanchard's data corrected for pupil area from his own measurements, and for relative effectiveness of pupil area from the work of Stiles and Crawford. The curves through the points are all the 2, 2 function in figure 2, corresponding to equation (7). As in figure 1, the data for yellow and red have been displaced upward 0.5 and 1 log unit, respectively, and the data for green and blue downward 0.5 and 1 log unit, respectively.

Blanchard's data corrected in this manner are shown in figure 3. The high intensity sections, representing cone vision, contain too few points (except for white) for a critical choice among the four curves of figure 2. The $m = n = 1$ curve is definitely excluded; and since for intensity discrimination the 2, 2 curve gave the best fit, the 2, 2 curve has been drawn through the present cone data as well. The fit of the data by the theoretical curve is obvious. The rod data at low intensities are numerous enough

to exclude the 1, 1 formulation. Flicker measurements[10] would encourage the use of the 2, 1 curve for the rods. Though this is not excluded here, the 2, 2 fits distinctly better; its curve is therefore drawn through the rod data also.

One must treat this choice of exponents as distinctly preliminary. The data are for one person only and have had to be corrected; additional measurements made with a fixed pupil are necessary before serious attention can be paid to the specific values of m and n. However, the measurements are good enough to show their agreement with theoretical expectation, especially since their precise form could not be predicted intuitively.

Summary.—Blanchard's measurements of the instantaneous visual threshold following light adaptation to different intensities and colors are shown to fit directly into the pattern of the duplicity theory which separates vertebrate vision into cone and rod functions. Moreover, the numerical values of the measurements and the precise form of their relations are found to conform to the mathematical predictions from a reversible photochemical system previously used for describing other properties of vision and photoreception.

[1] H. Aubert, *Physiologie der Netzhaut*, Breslau, E. Morgenstern, pp. 394 (1865).

[2] A. Koenig and E. Brodhun, *Sitzungsber. Akad. Wissensch.*, Berlin, **917** (1888).

[3] J. Blanchard, *Phys. Rev.*, Series 2, **11**, 81 (1918).

[4] S. Hecht, *Proc. Nat. Acad. Sci.*, **20**, 644 (1934).

[5] M. Schultze, *Arch. mikr. Anat.*, **2**, 175 (1866); H. Parinaud, *Compt. rend. Acad. Sci.*, Paris, **101**, 821 (1885); J. von Kries, in A. Bethe, G. von Bergmann, G. Embden, A. Ellinger, *Handbuch der normalen und pathologischen Physiologie*, J. Springer, **13**, pt. 1, 678 (1929); S. Hecht, *Physiol. Rev.*, **17**, 239 (1937).

[6] A. Kohlrausch, *Arch. ges. Physiol.*, **196**, 113 (1922).

[7] S. Hecht and S. Shlaer, *J. Gen. Physiol.*, **19**, 965 (1936).

[8] W. S. Stiles and B. H. Crawford, *Proc. Roy. Soc. London*, Series B, **112**, 428 (1933).

[9] L. T. Troland, *Trans. Ill. Eng. Soc.*, 947 (1916).

[10] S. Hecht and E. L. Smith, *J. Gen. Physiol.*, **19**, 979 (1936).

THE GROUPS OF ORDER p^m WHICH HAVE $m-1$ INDEPENDENT GENERATORS

BY G. A. MILLER

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF ILLINOIS

Communicated February 27, 1937

When a group G of order p^m, p being a prime number, has m independent generators it is abelian and of type 1^m and hence its properties are well known, but when it has $m-1$ independent generators a number of different types of G's present themselves. The case when p is odd is much simpler than the special case when $p=2$ but these two cases have some common properties. In both of them the pth powers of all the operators of G are contained in a subgroup of order p but in the former case these powers may be the identity for all the operators of a given group, while in the latter case every G involves operators of order 4, and hence the powers of its operators generate the given subgroup of order p. This subgroup always includes the commutator subgroup of G but in both cases the latter subgroup may be the identity. There is obviously one and only one such G for every value of p and for every value of m greater than 1. Since the direct product of any G which comes under the heading of this article and any abelian G of type 1^k also comes thereunder, such direct products will be excluded in what follows unless the contrary is stated. When G is abelian there is one and only one such group for a given value of p which is not a direct product.

If s_1 is a non-invariant operator of G all the operators of G which are commutative with s_1 constitute a subgroup of index p under G, and if s_2 is an operator of G which does not appear in this sub-group, then G involves a subgroup of index p^2 composed of all its operators which are commutative with both of the operators s_1, s_2. This subgroup includes the commutator subgroup of G and hence it is invariant under G. The corresponding quotient group is abelian and of type 1^2. There are $p-1$ co-sets corresponding to operators of this quotient group which are composed of all the operators of G which are commutative with s_1 but not with s_2 and $p-1$ other such co-sets composed of all the operators of G which are commutative with s_2 but not with s_1 while the remaining such co-sets are composed of operators which are not commutative with either one of the two operators s_1, s_2. The subgroup composed of all the operators of G which are commutative with both of the operators s_1, s_2 involves the same number of operators which are invariant under it as G does. If it is non-abelian we repeat the given operations until we arrive at an abelian subgroup by this process.

In the case when $p>2$ the operators of order p contained in G together

with the identity constitute a group but this is not necessarily the case when $p = 2$. Hence we shall confine our attention in what follows, except the last paragraph, to the case when p is odd. Since it is assumed that G is not a direct product and involves no operator of order p^2 every such group contains only p invariant operators. The subgroup composed of all of the operators of G which are commutative with both s_1 and s_2 contains also only p invariant operators but the subgroup of index p under G composed of all its operators which are commutative with one of its non-invariant operators contains p^2 invariant operators. Hence all these groups can be constructed in the following manner:

Start with the abelian group of type $1^{(m+1)/2}$ and extend it by an operator of order p which transforms it into itself and gives rise to a commutator of order p. We thus obtain a group whose central is of type $1^{(m-1)/2}$. When $m > 3$ we adjoin to this group another operator of order p which transforms it into itself and gives rise to the same commutator subgroup with respect to its central. The central of the resulting group is of type $1^{(m-1)/2}$. When $m > 5$ we adjoin an additional operator to this group which transforms it into itself and gives rise to the same commutator subgroup with respect to its central, etc. This process may be continued until we arrive at G, whose central is of order p. It is obvious that m may be an arbitrary odd positive integer which exceeds unity. Including direct products it therefore results that *the number of the distinct groups of order* p^m, $p > 2$ *which have* $m - 1$ *independent generators and involve no operator of order* p^2 *is* $(m-1)/2$ *when* m *is odd and* $(m-2)/2$ *when* m *is even.*

When G involves operators of order p^2, p being odd, the group generated by two non-commutative operators of this order is of order p^3 and involves a subgroup of order p^2 composed of its operators of order p together with the identity. This subgroup involves operators of order p which are non-commutative with each of its operators of order p^2. All the operators of G which are commutative with every operator of this subgroup constitute a subgroup of index p^2 under G. If this subgroup involves a non-invariant operator of order p^2 it also involves an operator of order p which is non-commutative with it and a subgroup of index p^2 under it composed of all its operators which are commutative with each of these two operators. By repeating this process we finally arrive at an abelian group which will be either of type 1^k or of type $2, 1^k$. Hence two cases present themselves. In one of these G involves invariant operators of order p^2 while in the other it does not have this property. In the former case G may be abelian but in the latter case it is necessarily non-abelian.

When G does not involve an invariant operator of order p^2, all the possible groups of order p^m can be constructed as follows: Start with the

abelian group of order $p^{(m+1)/2}$ which contains no operator of order p^2 just as in the case when G involves no operator of order p^2. In the present case the group may be extended arbitrarily by an operator of order p or by an operator of order p^2 except that at least one of the $(m-1)/2$ extending operators must be of order p^2. Hence there result $m-2$ distinct groups since the extending operators can always be so selected that they have different orders except in one case. When G involves an invariant operator of order p^2, all the possible groups can be constructed by starting with the abelian subgroup of type 2, $1^{(m-2)/2}$. Since the extending operator can be selected so as to have either of two different orders in each of the $(m-2)/2$ possible cases, the number of the possible non-abelian G's in this case is again $m-2$. As there is also one possible abelian group it results that, *including direct products there are* $m-1$ *groups of order* p^m *which have the property that each of them involves invariant operators of order* p^2 *and contains exactly* $m-1$ *independent generators,* m *being even. When* m *is odd this number is* $m-2$.

In the special case when $p=2$ a necessary and sufficient condition that a group of order p^m has $m-1$ independent generators is that the squares of all its operators generate the subgroup of order 2. It is known that the number of these groups of order 2^m is $3(m-1)/2$ when m is odd and $(3m-4)/2$ when m is even.[1] Although these groups are somewhat more complex than those relating to the more general case when p is odd they seem to require no consideration here since their fundamental properties were determined in the article to which we referred at the close of the preceding sentence and in those referred to therein.

[1] G. A. Miller, these PROCEEDINGS, **22**, 112 (1936).

ON THE REGULAR REPRESENTATIONS OF ALGEBRAS

BY R. BRAUER AND C. NESBITT

UNIVERSITY OF TORONTO

Communicated February 24, 1937

The regular representations play an important rôle in the work of *Molien, Cartan* and *Frobenius*[1] in the theory of hypercomplex numbers. More recently, the theory of groups of linear transformations has been extended and new concepts have been introduced. Our first aim was to study the regular representations of an algebra with regard to these new ideas.

We consider an associative algebra A over a field F, assuming that A has a unit element. Since we shall be concerned with the absolutely irreducible constituents of the regular representations of A, we may assume without restriction that F is algebraically closed.

Let $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ be a basis of A. For every α in A, we have equations

$$\epsilon_\kappa \alpha = \sum_\lambda r_{\kappa\lambda} \epsilon_\lambda, \tag{1}$$

$$\alpha \epsilon_\kappa = \sum_\lambda s_{\lambda\kappa} \epsilon_\lambda \tag{2}$$

where the coefficients $r_{\kappa\lambda}$ and $s_{\kappa\lambda}$ lie in F. We then obtain two representations $\mathfrak{R}$ and $\mathfrak{S}$ of A by associating the matrices $R = (r_{\kappa\lambda})$, $S = (s_{\kappa\lambda})$ with α. These are the two *regular representations*. The decomposition of $\mathfrak{R}$ and of $\mathfrak{S}$ is our first concern.

Let $\mathfrak{A}$ be a representation of an algebra A by linear transformations of a vector space V. Then, $\mathfrak{A}$ is *decomposable* if V is the direct sum of two vector spaces V_1 and V_2 both invariant under $\mathfrak{A}$. Adapting the coördinate system in V to this decomposition, we obtain $\mathfrak{A}$ in the form

$$\mathfrak{A} = \begin{pmatrix} \mathfrak{A}_1 & 0 \\ 0 & \mathfrak{A}_2 \end{pmatrix} \tag{3}$$

where $\mathfrak{A}_1$ operates in V_1 and $\mathfrak{A}_2$ in V_2. The representations $\mathfrak{A}_1$ and $\mathfrak{A}_2$ may still be decomposable. Writing them in the same form as $\mathfrak{A}$ in (3) and continuing in this manner we finally obtain the splitting of $\mathfrak{A}$ into *indecomposable constituents*. These are uniquely determined[2] if equivalent representations are considered as equal. The indecomposable constituents may still be reducible.[3] We get, of course, the irreducible constituents of $\mathfrak{A}$ if we break up every indecomposable constituent into its irreducible parts.

Let us denote by $\mathfrak{U}_1, \mathfrak{U}_2, \ldots, \mathfrak{U}_k$ the non-equivalent indecomposable constituents of $\mathfrak{R}$ and by $\mathfrak{V}_1, \mathfrak{V}_2, \ldots, \mathfrak{V}_l$ the non-equivalent indecomposable constituents of $\mathfrak{S}$. Let $\mathfrak{F}_1, \mathfrak{F}_2, \ldots, \mathfrak{F}_m$ be the totality of all non-equivalent irreducible representations of A. Combining the powerful method used by *Cartan*[1] with new methods we obtain the following results which we state here without proof:

The numbers k, l and m of the $\mathfrak{U}_\kappa$, the $\mathfrak{V}_\lambda$ and the $\mathfrak{F}_\mu$, respectively, are equal. *We may arrange the* $\mathfrak{V}_\lambda$ *and the* $\mathfrak{F}_\mu$ *in such a manner that* $\mathfrak{V}_\lambda$ *contains* $\mathfrak{F}_\lambda$ *as first irreducible constituent* ($\lambda = 1, 2, \ldots, k$), i.e.,

$$\mathfrak{V}_\lambda = \begin{pmatrix} \mathfrak{F}_\lambda & 0 \\ * & \widetilde{\mathfrak{V}}_\lambda \end{pmatrix} \tag{4}$$

where $\widetilde{\mathfrak{V}}_\lambda$ is a (reducible or irreducible) representation of A, and the asterisk stands for terms about which we are not concerned. In (4) we include

the case that $\mathfrak{B}_\lambda$ itself is equal to $\mathfrak{F}_\lambda$. Moreover, if we have a splitting of $\mathfrak{B}_\lambda$ into constituents

$$\mathfrak{B}_\lambda = \begin{pmatrix} \mathfrak{H} & 0 \\ * & \mathfrak{K} \end{pmatrix}$$

and $\mathfrak{H}$ is completely reducible, then $\mathfrak{H}$ must be equivalent to $\mathfrak{F}_\lambda$. In other words: If we split $\mathfrak{B}_\lambda$ into largest completely reducible constituents in the sense of *A. Loewy*,[4] the first constituent of this form is $\mathfrak{F}_\lambda$ itself. This shows, in particular, that the one-to-one correspondence between $\mathfrak{B}_\lambda$ and $\mathfrak{F}_\lambda$ established in (4) is uniquely determined.

We obtain a corresponding result for $\mathfrak{R}$ if we consider the last irreducible constituent of $\mathfrak{U}_\lambda$ instead of the first. This shows: *We can numerate the* $\mathfrak{U}_\lambda$ *in such a way that* $\mathfrak{F}_\lambda$ *is the last constituent in* $\mathfrak{U}_\lambda$,

$$\mathfrak{U}_\lambda = \begin{pmatrix} * & 0 \\ * & \mathfrak{F}_\lambda \end{pmatrix}. \tag{5}$$

Here $\mathfrak{U}_\lambda$ is uniquely determined by $\mathfrak{F}_\lambda$. It is not possible to find another splitting of $\mathfrak{U}_\lambda$ of the form (5), in which there stands another completely reducible representation at the place of $\mathfrak{F}_\lambda$. In what follows the numeration of the $\mathfrak{U}_\lambda$ and $\mathfrak{B}$ will be as in (4) and (5).

We say that an indecomposable constituent $\mathfrak{U}$ of a representation $\mathfrak{A}$ has the *multiplicity* f, if f of the indecomposable constituents of $\mathfrak{A}$ are equivalent to $\mathfrak{U}$. Similarly, we may define the multiplicity of an irreducible constituent of $\mathfrak{A}$. We denote the degree of $\mathfrak{F}_\lambda$ by f_λ, that of $\mathfrak{U}_\lambda$ by u_λ, and that of $\mathfrak{B}_\lambda$ by v_λ. Then we can show that $\mathfrak{U}_\lambda$ has the multiplicity f_λ as an indecomposable constituent of $\mathfrak{R}$, and $\mathfrak{B}_\lambda$ has the same multiplicity f_λ in $\mathfrak{S}$. On the other hand, $\mathfrak{F}_\lambda$ appears as an irreducible constituent of multiplicity v_λ in $\mathfrak{R}$ and of multiplicity u_λ in $\mathfrak{S}$.[5] It is easy to give examples of algebras for which u_λ is different from v_λ.

We shall denote by $c_{\kappa\lambda}$ the multiplicity of $\mathfrak{F}_\lambda$ as irreducible constituent of $\mathfrak{B}_\kappa$. The number $c_{\lambda\kappa}$ then gives the multiplicity of $\mathfrak{F}_\lambda$ as irreducible constituent of $\mathfrak{B}_\kappa$. Considering the degrees of the different representations we obtain the relations,

$$u_\kappa = \sum_\lambda c_{\kappa\lambda} f_\lambda \,, \quad v_\kappa = \sum_\lambda c_{\lambda\kappa} f_\lambda \,, \tag{6}$$

$$n = \sum_\xi u_\xi f_\xi = \sum_\xi v_\xi f_\xi = \sum_{\kappa,\lambda} c_{\kappa\lambda} f_\kappa f_\lambda = \sum_{\kappa,\lambda} c_{\lambda\kappa} f_\kappa f_\lambda. \tag{7}$$

Another property of the numbers $c_{\kappa\lambda}$ is connected with the matrices intertwining the indecomposable constituents. Let $\mathfrak{B}$ and $\mathfrak{C}$ be any two representations of A which associate the matrices $B(\alpha)$ and $C(\alpha)$ with the

element α of A. The representation $\mathfrak{B}$ is *intertwined* with $\mathfrak{C}$, if there exists a matrix $P \neq 0$, independent of α, for which

$$B(\alpha)P = PC(\alpha) \qquad \text{for all } \alpha \text{ in A} \tag{8}$$

holds.[6] We say that $\mathfrak{B}$ is *h times intertwined* with $\mathfrak{C}$ if there exist exactly h linearly independent matrices P satisfying (8). We then set $h = I(\mathfrak{B},\mathfrak{C})$. If $\mathfrak{B}$ is a representation containing $\mathfrak{F}_\kappa$ exactly h times as an irreducible constituent, the relations hold,

$$h = I(\mathfrak{U}_\kappa, \mathfrak{B}) = I(\mathfrak{B}, \mathfrak{B}_\kappa). \tag{9}$$

This implies the following characterization of the $c_{\kappa\lambda}$:

$$c_{\kappa\lambda} = I(\mathfrak{U}_\kappa, \mathfrak{B}_\lambda) = I(\mathfrak{U}_\lambda, \mathfrak{U}_\kappa) = I(\mathfrak{B}_\lambda, \mathfrak{B}_\kappa). \tag{10}$$

A third characterization of the $c_{\kappa\lambda}$ can be obtained from Cartan's ideas.

The quotient algebra A/N, where N is the radical of A is a sum of simple algebras H_λ. Let the elements η_λ (mod N) give the unit elements of these H_λ. The number l of elements η_λ is equal to the number k of the $\mathfrak{F}_\kappa$ and we may numerate the η_κ in such a way that η_κ is represented by the unit matrix in $\mathfrak{F}_\kappa$, and by zero in the other $\mathfrak{F}_\mu$. We may choose the η_λ so that

$$\eta_\kappa^2 = \eta_\kappa, \qquad \eta_\kappa\eta_\lambda = 0 \text{ for } \kappa \neq \lambda.$$

Then $c_{\kappa\lambda}f_\kappa f_\lambda$ *gives the number of linearly independent elements* α *in* A *for which* $\eta_\kappa\alpha\eta_\lambda = \alpha$ *holds.* Cartan's results concerning the determinants of the regular representations are contained in those given above.

A fourth characterization of the $c_{\kappa\lambda}$ has been given by Frobenius.[7] We put

$$\alpha = x_1\epsilon_1 + x_2\epsilon_2 + \ldots + x_n\epsilon_n, \quad \beta = y_1\epsilon_1 + y_2\epsilon_2 + \ldots + y_n\epsilon_n$$

and denote by $R(\alpha)$, $R(\beta)$, $S(\alpha)$, $S(\beta)$ the matrices representing α and β in $\mathfrak{R}$ and $\mathfrak{S}$, respectively. The $c_{\kappa\lambda}$ then appear as exponents of the irreducible factors of the polynomial

$$\det (S(\alpha) + R(\beta)') = \varphi(x_1, \ldots, x_n, y_1, \ldots, y_n)$$

in $x_1, \ldots, x_n, y_1, \ldots, y_n$.

We consider now algebras A for which the two regular respresentations are equivalent. A necessary and sufficient condition for this equivalence has been given by *Frobenius*,[1] namely that the parastrophic determinant should not identically vanish. We shall denote such an A as a *Frobenius algebra*. If we consider A as an n-dimensional vector space, we can express the characteristic condition for Frobenius algebras in the following form: Not every hyperplane in A contains a right ideal.

In case of a Frobenius algebra we certainly have $u_\lambda = v_\lambda$. The totality $\mathfrak{U}_1, \mathfrak{U}_2, \ldots, \mathfrak{U}_k$ of indecomposable constituents of $\mathfrak{R}$ coincides with the

totality of indecomposable constituents of $\mathfrak{S}$. It is, however, possible that $\mathfrak{U}_\kappa$ is not equivalent to $\mathfrak{B}_\kappa$. Moreover, there exist examples where $c_{\kappa\lambda}$ is different from $c_{\lambda\kappa}$. This led us to seek a sub-class of the Frobenius algebras in which $\mathfrak{U}_\kappa$ and $\mathfrak{B}_\kappa$ would be equivalent.

We will call an algebra *symmetric*, if the following condition is satisfied: *There exists a hyperplane in* A *which contains all commutator elements* $\alpha\beta - \beta\alpha$, *but does not contain a right ideal.* Every such algebra is, of course, a Frobenius algebra. The semi-simple algebras belong to the symmetric algebras, and so do the group rings of finite groups, even in the case when the latter are not semi-simple, i.e., the case where the characteristic of the field is a prime dividing the order of the group.

A necessary and sufficient condition for an algebra A to be symmetric is that there exists a symmetric matrix P transforming $\mathfrak{N}$ into $\mathfrak{S}$.

In the case of symmetric algebras, we can prove that $\mathfrak{U}_\kappa$ and $\mathfrak{B}_\kappa$ are equivalent. This implies that *the first and the last irreducible constituent of* $\mathfrak{U}_\kappa$ *are the same.* It also follows immediately that $c_{\kappa\lambda} = c_{\lambda\kappa}$ for all κ, λ. In particular, these results hold for the representations of finite groups in Galois fields.

[1] Th. Molien, *Math. Ann.*, **41**, 83 (1893); E. Cartan, *Annales de Toulouse*, **12**, B1 (1898); G. Frobenius, *Sitzungsber. Preuss. Akad. Wiss.*, **1903**, 401 and 634.

[2] W. Krull, *Math. Zeit.*, **23**, 161 (1925). Cf. also R. Brauer–I. Schur, *Sitzungsber. Preuss. Akad. Wiss.*, **1930**, 209.

[3] See, for instance, H. Weyl, *The Theory of Groups and Quantum Mechanics*, London (1931), pp. 121–122 for the definition of irreducibility and equivalence.

[4] *Trans. Amer. Math. Soc.*, **6**, 504 (1905).

[5] Cf. R. Brauer, *Actualités scientifiques et industrielles*, No. 195 (1935). Theorem II.

[6] Cf. I. Schur, *Sitzungsber. Preuss. Akad. Wiss.*, **1905**, 406.

[7] Frobenius, § 11 (3).

SIMPLE LIE ALGEBRAS OF TYPE A

BY N. JACOBSON[1]

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF CHICAGO

Communicated March 8, 1937

If $\mathfrak{A}$ is an associative algebra over Φ it becomes a Lie algebra[2] when $[a, b]$ is defined as $ab - ba$ where ab is the associative product originally given in $\mathfrak{A}$. If in addition $\mathfrak{A}$ is self-reciprocal, i.e., there is defined a correspondence $a \longrightarrow a^J \epsilon \mathfrak{A}$ such that

$$(a + b)^J = a^J + b^J \quad (\alpha a)^J + \alpha a^J \quad (ab)^J = b^J a^J, \qquad \alpha \epsilon \Phi$$

then the elements a such that $a^J = -a$ are called *J-skew* and their totality

$\mathfrak{S}_J$ is easily seen to be a Lie subalgebra of $\mathfrak{A}$. If $a^{J^2} = a$ for all a, J is called an *involutorial anti-automorphism* (i.a.a.). Then $K = S^{-1}JS$ is also an i.a.a. for S an arbitrary automorphism of the associative algebra $\mathfrak{A}$ over Φ and we call K *cogredient* to J. It is easily seen that $\mathfrak{S}_K$ consists of the elements a^S where a varies in $\mathfrak{S}_J$ and so $\mathfrak{S}_J$ and $\mathfrak{S}_K$ are isomorphic. Hence the derived algebras $\mathfrak{S}_J'$ and $\mathfrak{S}_K'$ are isomorphic also. If $J = K$, i.e., $SJ = JS$ then S is an automorphism of $\mathfrak{A}$ inducing an automorphism in the Lie algebras $\mathfrak{S}_J$ and $\mathfrak{S}_J'$.

In a paper[3] which will appear shortly in the *Annals of Mathematics* I discussed the Lie algebras $\mathfrak{S}_J$ obtainable in this way from a normal simple algebra $\mathfrak{A}$ over Φ of characteristic 0 and showed their connection with the theory of abstract Lie algebras of types B, C and D. The present note will give an outline of an extension of these results to involutorial simple algebras of second kind as defined by Albert.[4]

Following Landherr[5] we say that a Lie algebra $\mathfrak{B}$ over Φ of characteristic 0 has *type* A if $\mathfrak{B}_\Omega \cong \Omega_n'$ where Ω is the algebraic closure of Φ and Ω_n the algebra of all n-rowed matrices with coördinates in Ω. $\mathfrak{B}$ has *type* A_{I} or A_{II} according as the coefficients of the characteristic equations of the matrices representing the elements of $\mathfrak{B}$ in Ω_n all belong to Φ or not. In either case $\mathfrak{B}$ is simple since Ω_n' is.

Now suppose $\mathfrak{A}$ is simple of order $2n^2$ over Φ of characteristic 0 with $\mathrm{P} = \Phi(q)$, $q^2 = \mu$ in Φ as centrum and J an i.a.a. of second kind in $\mathfrak{A}$, i.e., $\xi^J = \bar{\xi} = \alpha - \beta q$ for $\xi = \alpha + \beta q$ in P. We prove

THEOREM 1. *If* $\mathrm{n} > 2$ $\mathfrak{S}_J'$ *is a Lie algebra of type* A_{II}.

An important lemma used in the proof is that the enveloping algebra over Φ of $\mathfrak{S}_J'$ is $\mathfrak{A}$. The classification of the Lie algebras $\mathfrak{S}_J'$ and their automorphisms are given by

THEOREM 2. *If* $\mathfrak{A}_1$ *and* $\mathfrak{A}_2$ *are involutorial simple algebras of order* $2\mathrm{n}^2$ *over* Φ, $\mathrm{n} > 2$, *and* J_1 *and* J_2 *are i.a.a. of second kind in* $\mathfrak{A}_1$ *and* $\mathfrak{A}_2$, *respectively, such that* $\mathfrak{S}_{J_1}' \cong \mathfrak{S}_{J_2}'$ *then* $\mathfrak{A}_1 \cong \mathfrak{A}_2$ *and when* $\mathfrak{A}_1$ *and* $\mathfrak{A}_2$ *are identified* J_1 *and* J_2 *are cogredient.*

THEOREM 3. *If* $\mathfrak{A}$ *is involutorial simple of order* $2\mathrm{n}^2$, $\mathrm{n} > 2$, *over* Φ *and* J *an i.a.a. of second kind then any automorphism in* $\mathfrak{S}_J'$ *may be extended to a unique automorphism of the associative algebra* $\mathfrak{A}$.

The extended automorphism commutes with J. The inner automorphism $a \longrightarrow g^{-1}ag$ commutative with J are given by *J-orthogonal* elements, i.e., $gg^J = \gamma \neq 0$ in Φ. These form an invariant subgroup of index 1 or 2 in the group of all automorphisms of $\mathfrak{A}$ commutative with J. It follows that if $\mathfrak{g}_J$ is the group of J-orthogonal elements and $\mathfrak{d}$ the set of multiples $\delta 1$, $\delta \neq 0$ in P then $\mathfrak{g}_J/\mathfrak{d}$ is isomorphic either to the complete group of automorphisms of $\mathfrak{S}_J'$ or to an invariant subgroup of index 2 of this group.

As a converse of Theorem 1 we have

THEOREM 4. *If $\mathfrak{B}$ is a Lie algebra of type* A_{II} *then* $n > 2$ *and* $\mathfrak{B} \cong \mathfrak{S}_J'$ *where* J *is an i.a.a. of second kind in a simple algebra* $\mathfrak{A}$.

Thus a complete classification of Lie algebras of type A_{II} depends on the classification of simple associative algebras of second kind and of i.a.a.'s relative to cogredience in these algebras. It can be shown that the latter problem is essentially one of ordinary cogredience of hermitian matrices with coördinates in an involutorial division algebra of second kind. These questions are largely arithmetic in character and depend on the structure of Φ.

A detailed account of this theory will appear in another journal.

[1] NATIONAL RESEARCH FELLOW.

[2] For definitions of the principal concepts in the theory of Lie algebras, see N. Jacobson, "Rational Methods in the Theory of Lie Algebras," *Ann. Math.*, **36**, 875–881 (1935).

[3] "A Class of Normal Simple Lie Algebras of Characteristic 0." An abstract of this paper appeared in the *Bull. Am. Math. Soc.*, **43**, Jan. (1937).

[4] A. A. Albert, "Involutorial Simple Algebras and Riemann Matrices," *Ann. Math.*, **36**, 894 (1935).

[5] W. Landherr, "Ueber einfache Liesche Ringe," *Hamb. Abhandlungen*, **11**, 50 (1935). $\mathfrak{B}_\Omega$ denotes the algebra obtained from $\mathfrak{B}$ by extending the field Φ to Ω. The Lie algebras of type A_1 were determined completely by Landherr.

THE ITERATED STIELTJES TRANSFORM

BY D. V. WIDDER

DEPARTMENT OF MATHEMATICS, HARVARD UNIVERSITY

Communicated March 12, 1937

In a paper by the present author, which has not yet been published, the inversion of the Stieltjes transform

$$f(x) = \int_0^\infty \frac{d\alpha(t)}{x+t} \tag{1}$$

$$f(x) = \int_0^\infty \frac{\varphi(t)}{x+t}\,dt \tag{2}$$

was discussed. It was found that the integral (2) could be inverted by a linear differential operator of infinite order. This operator is one which has for its fundamental solutions all integral powers of x,

$$x^n \qquad (n = 0, \pm 1, \pm 2, \ldots).$$

More precisely, if we set

$$L_{k,t}[f(x)] = (-1)^{k-1} C_k \frac{d^k}{dt^k} t^{2k-1} \frac{d^{k-1}}{dt^{k-1}} f(t),$$

$$C_k = \frac{1}{k!(k-2)!} \quad (k = 2, 3, \ldots),$$

$$C_1 = 1,$$

then, for the integral (2) we have

$$\lim_{k \to \infty} L_{k,t}[f(x)] = \varphi(t)$$

for almost all positive values of t. For the integral (1) we have

$$\lim_{k \to \infty} \int_0^t L_{k,u}[f(x)]du = \frac{\alpha(t+)+\alpha(t-)}{2} - \alpha(0+).$$

By the iterated Stieltjes transform we mean

$$f(x) = \int_0^\infty \frac{\log \frac{x}{t}}{x-t} d\alpha(t), \tag{3}$$

$$f(x) = \int_0^\infty \frac{\log \frac{x}{t}}{x-t} \varphi(t)dt, \tag{4}$$

for these integrals are clearly obtained by formally iterating (1) and (2), respectively. It is to be noted that the kernel now becomes infinite at $t = 0$, so that we must understand by (3) the Cauchy value of the integral.

It is natural to suppose that (4), for example, can be inverted by two successive applications of the infinite differential operator defined above. However, this would involve two limit processes. We find that it is possible to introduce a new inversion operator involving a single limit process. It is again a linear differential operator of infinite order. Its fundamental solutions are

$$x^n \qquad (n = 0, \pm 1, \pm 2, \ldots)$$

$$x^n \log x \qquad (n = 0, \pm 1, \pm 2, \ldots).$$

More precisely, set

$${}_kL^*_{k,t}[f(x)] = C_k^2 \frac{d^k}{dt^k} t^{2k-1} \frac{d^{2k-1}}{dt^{2k-1}} t^{2k-1} \frac{d^{k-1}}{dt^{k-1}} f(t).$$

We can show that if $f(x)$ has the representation (4), then

$$L^*_{k,t}[f(x)] = \int_0^\infty F_k(u, t)\varphi(u)du, \tag{5}$$

where

$$F_k(u, t) = (2k-1)!(2k-1)!\, C_k^{2k-1} t^{k} u^{k} \int_0^\infty \frac{y^{2k-1} dy}{(t+y)^{2k}(u+y)^{2k}}.$$

An investigation of the behavior of the integral (5) as k becomes infinite shows that

$$\lim_{k \to \infty} L^*_{k,t}[f(x)] = \varphi(t)$$

for almost all positive values of t. We can also show that for the integral (3)

$$\lim_{k \to \infty} \int_0^t L_{k,u}[f(x)]du = \frac{\alpha(t+) + \alpha(t-)}{2} \quad (t > 0).$$

This inversion operator enables us to discuss the possibility of representing functions $f(x)$ in the forms (3) and (4). For example, one can show that $f(x)$ has the form (4) with $\varphi(t)$ belonging to the class L^p $(p > 1)$ if and only if there exists a constant M such that

$$\int_0^\infty \left| L^*_{k,t}[f(x)] \right|^p dt < M \quad (k = 1, 2, \ldots),$$

$$f(x) = o\left(\frac{1}{x}\right). \qquad (x \longrightarrow 0).$$

In a later paper by R. P. Boas and the author, detailed account of such representation theorems will be given.

METRIC METHODS IN CALCULUS OF VARIATIONS

By Karl Menger

University of Notre Dame

Communicated March 5, 1937

1. In each problem of calculus of variations there are four fundamental data: (I) A *point-space* S or a domain of a point-space. (II) A *function* φ from which there is derived a functional in a class $\mathfrak{C}$ of entities of S whose dimensions are $\geqq 1$. With each element C of $\mathfrak{C}$ there is associated a number, say $\lambda_\varphi(C)$, by a process of integration of φ along C. The class $\mathfrak{C}$ is a topological space in which a metric may be introduced. (III) A sub-class $\mathfrak{C}^*$ of $\mathfrak{C}$, the *class of admissible entities*. (IV) A sub-class $\mathfrak{C}^*_0$ of $\mathfrak{C}^*$. A main problem is to find conditions under which $\mathfrak{C}^*_0$ contains elements minimizing $\lambda_\varphi(C)$ or stationary for $\lambda_\varphi(C)$ with respect to all elements of $\mathfrak{C}^*$. The solution of this problem is the more general the fewer re-

strictions are imposed on S, φ and $\mathfrak{C}^*$, and the more restrictions are imposed on $\mathfrak{C}^*_0$. In fact, one of the main tendencies in calculus of variations has been to find solutions of this problem more and more general in the mentioned sense.[1]

2. The existence of minimizing elements was proved by Tonelli essentially by an implicit application of the following slight generalization of Baire's well-known theorem that a lower semicontinuous (l.s.c.) function assumes its minimum on each compact set: Let $\mathfrak{C}$ be a topological space, $\lambda(C)$ a real function (not necessarily finite) defined for all elements of $\mathfrak{C}$, $\lambda_\varphi(C)$ a finite function defined on the set $\mathfrak{C}_f$ of all elements C of $\mathfrak{C}$ for which $\lambda(C)$ is finite. For each finite number ρ the following conditions may be satisfied: (1) the set $\mathfrak{C}_\rho$ of all C with $\lambda(C) \leqq \rho$ is compact; (2) $\lambda_\varphi(C)$ is l.s.c. on $\mathfrak{C}_\rho$; (3) $\lambda(C)$ is bounded above on the set of all C with $\lambda_\varphi(C) \leqq \rho$. *Then on each closed sub-set $\mathfrak{C}^*$ of $\mathfrak{C}_f$ the function $\lambda_\varphi(C)$ assumes its minimum.* Suppose now that λ_φ satisfies in addition to (1) and (2) the condition (2′) $\lim C_n = C$ and $\lim \lambda(C_n) = +\infty$ always implies that $\lim \lambda_\varphi(C_n) = +\infty$. Then[2] condition (3) is equivalent to the following condition (3′): For each ρ the set of all C with $\lambda_\varphi(C) \leqq \rho$ is compact. If λ_φ besides (1) and (2) satisfies the condition (4) $\lambda(C) = +\infty$ always implies $\lambda_\varphi(C) = +\infty$, *then on each closed sub-set $\mathfrak{C}^*$ of $\mathfrak{C}$* (not only of $\mathfrak{C}_f$) $\lambda_\varphi(C)$ *attains its minimum, and for an element of $\mathfrak{C}_f$* (unless for each element C of $\mathfrak{C}^*$ we have $\lambda_\varphi(C) = +\infty$). Tonelli used the first-mentioned theorem in the case that $\mathfrak{C}$ is the class of all continuous curves of a space S, that $\lambda_\varphi(C)$ is a curve-integral along C, and that $\lambda(C)$ is the length of C, thus $\mathfrak{C}_f$ the set of all rectifiable curves. Morse has deduced an extensive theory concerning the existence and the number of stationary curves from hypotheses (2), (2′) and related assumptions, by profound topological methods.

3. The main part of Tonelli's theory is a proof, by methods of the theory of real functions, that under the following assumptions on S, φ and $\mathfrak{C}^*$ the curve-integral $\lambda_\varphi(C)$ along C actually satisfies the conditions (2) and (2′) of §2 with respect to the length $\lambda(C)$: I. S is *a domain of n-dimensional euclidean space.* Each point x of S is defined by n real numbers $\xi_i (i = 1, \ldots, n)$. II. The function of point and direction, thus a function $\varphi(\xi_i, \xi_i')$ depending on $2n$ real variables ξ_i and ξ_i' (on the ξ_i' in a positive homogeneous way) is *everywhere continuous* in the $2n$ variables (or even differentiable with respect to the ξ_i') and *"quasiregular" in each point* ξ_i, i.e., Legendre's condition in the weak sense (with the $\geqq$ sign) holds if second derivatives of φ with respect to the ξ_i' exist, otherwise properties of convexity are supposed for the n-dimensional surface $\eta = \varphi(\xi_i, \xi_i')$ over the n-dimensional ξ_i'-space for each fixed point ξ_i. In order to guarantee condition (2′) of §2 Tonelli assumed φ, furthermore, as *"seminormal,"* i.e., for no point ξ_i may Legendre's expression vanish for all directions ξ_i'. The number $\lambda_\varphi(C)$ associated with the curve $\xi_i(t)$ ($t_0 \leqq$

$t \leqq t_1$) is $\int_{t_0}^{t_1} \varphi(\xi_i(t),\xi_i'(t))dt$. This number is determined if the n functions $\xi_i(t)$ are differentiable for all values of t with a few exceptions, or at most with the exception of a set of linear measure 0, after Hahn's introduction of Lebesgue's integral. According to Lebesgue they are certainly differentiable almost everywhere if C is rectifiable and if $\mathfrak{C}$ is the class of all rectifiable curves. III. Hence $\mathfrak{C}^*$ was assumed by Tonelli as *a class of rectifiable curves*. Methods of metric geometry[3] allow one to generalize these assumptions on S, φ and $\mathfrak{C}^*$ considerably.

4. *The space* S. We admit that S be a set of elements whatever ("points") such that with each pair of points q,r there is associated a number $\delta(q,r)$ satisfying the following conditions: (1) $\delta(q,r) = \delta(r,q) > 0$ if $q \neq r$, while $\delta(q,q) = 0$. (2) For each $\epsilon > 0$ there exists a $\delta > 0$ such that $\delta(q,q') + \delta(r,r') < \delta$ always implies $| \delta(p,q) - \delta(p',q') | < \epsilon$. (3) For each $\xi > 0$ there exists a $\delta > 0$ such that for each polygon (finite ordered set) $P = \{p_1, \ldots, p_m\}$ whose diameter is $< \delta$ we have $\lambda(P) > (1 - \xi)\mu(P)$, where $\lambda(P) = \sum_{i=1}^{m-1} \delta(p_i,p_{i+1})$ and $\mu(P) = \delta(p_1,p_m)$. The metric spaces in the sense of Fréchet are particular examples of our general spaces since the triangular inequality implies the conditions (2) and (3). If C is a continuous curve in our space S, i.e., a continuous image of a closed segment $[\alpha,\beta]$ of real numbers, then C has a length, that is to say there exists a number $\lambda(C)$, finite or $+\infty$, such that $\lim \lambda(P_n) = \lambda(C)$ for each sequence of sub-polygons $P_1,P_2, \ldots$ of the curve C, whose norms tend toward 0. Thereby we call a sub-polygon of C, each polygon $p(\gamma_1),p(\gamma_2), \ldots, p(\gamma_m)$ where $\alpha = \gamma_1 < \gamma_2 < \ldots < \gamma_{m-1} < \gamma_m = \beta$ and norm of P the largest of the $m - 1$ numbers $\gamma_i - \gamma_{i-1}$.

5. *The function* φ. There need not be defined directions in S. For this reason we suppose φ to be a function *of the points of* S *and the pairs of distinct points of* S. If such a function $\varphi(p;q,r)$ is given[4] we may derive from S a new space S_φ, called S *distorted by* φ, with the same points as S but with the distance $\delta_\varphi(q,r) = \varphi(q;q,r)\delta(q,r)$ if $q \neq r$ and $\delta_\varphi(q,q) = 0$. It is useful to define for each point p of S a space $S_{\varphi p}$ *tangential to* S_φ *in* p, consisting of the points of S_φ with the distance $\delta_{\varphi p}(q,r) = \varphi(p;q,r)\delta(q,r)$.

We call $\varphi(p;q,r)$ *l.s.c.* with respect to p in p_0 if for each $\epsilon > 0$ there exists a $\delta > 0$ such that $\delta(p,p_0) < \delta$ always implies $\varphi(p_0;q,r) - \varphi(p;q,r) < \epsilon$ for all points q and r $(\neq q)$. A measure for the deviation from lower semicontinuity in p_0 is the number $\sigma(p_0)$ defined as the limit for $\rho \longrightarrow 0$ of the upper bound of the numbers $\varphi(p_0;q,r) - \varphi(p;q,r)$ for all points p with $\delta(p,p_0) < \rho$ and all q,r.

We call $\varphi(p;q,r)$ *quasiregular in* p if in $S_{\varphi p}$ the triangular inequality $\delta_{\varphi p}(p,q) + \delta_{\varphi p}(q,r) \geqq \delta_{\varphi p}(p,r)$ holds for each pair of points q,r; then we also have the inequality $\lambda_{\varphi p}(P) \geqq \mu_{\varphi p}(P)$ for each polygon $P = \{p_1, \ldots, p_m\}$

with $p_1 = p$, where $\lambda_{\varphi p}(P) = \sum\delta_{\varphi p}(p_i,p_{i+1})$ and $\mu_{\varphi p}(P) = \delta_{\varphi p}(p_1,p_m)$. A measure for the deviation from quasiregularity in p is the number $\tau(p)$ defined as the upper bound of the numbers $[\lambda_{\varphi p}(P) - \mu_{\varphi p}(P)]:\left|\ \mu_{\varphi p}(P)\ \right|$ for all polygons with $p_1 = p$.

All we assume on φ is merely that it satisfies the following hypotheses with respect to each rectifiable curve C:

H_1. *There exists a neighborhood* U *of* C *such that* φ(p;q,r) *is bounded for all points* p,q,r *of* U.

H_2. *For each* $\sigma > 0$ *the set of all points* p *of* C *with* $\sigma(p) \geqq \sigma$ *has*[5] *the linear measure* 0, i.e., may be covered, for each $\delta > 0$, by a finite number of spheres whose diameters have a sum $< \delta$.

H_3. *For each* $\tau > 0$ *the set of all points of* C *with* $\tau(p) \geqq \tau$ *has the linear measure* 0.

H_4. *The upper bound of the numbers* τ(p) *for all points* p *of* C *is finite.*

H_5. *Each point* p_0 *for which we have* $\sigma(p_0) > 0$ *or which is point of accumulation of points* p *with* $\sigma(p) > 0$, *is contained in a neighborhood* U *such that* $\lambda_\varphi(P) \geqq 0$ *for all*[6] *closed polygons* $P = \{p_1, p_2, \ldots, p_m = p_1\}$ *of* U.

Under these hypotheses it can be proved that each rectifiable curve C has a finite length[7] $\lambda_\varphi(C)$ in S_φ. That is to say, there exists a finite number $\lambda_\varphi(C)$ such that for each sequence of sub-polygons $P_1, P_2, \ldots$ of the curve C whose norms tend toward 0 we have $\lim \lambda_\varphi(P_n) = \lambda_\varphi(C)$. For each finite number ρ this functional $\lambda_\varphi(C)$ is l.s.c. on the set of all curves of lengths $\leqq \rho$. Thus in the topological space $\mathfrak{C}$ of all curves of S the function $\lambda_\varphi(C)$ satisfies condition (2) of §2, and $\lambda(C)$ satisfies condition (1) if S is compact. To condition (3) of §2 there corresponds the hypothesis

H_6. *For each finite number* ρ *the lengths* λ(C) *of all curves* C *with* $\lambda_\varphi(C) \leqq \rho$ *are bounded.*

From the first theorem of §2 it follows: *If we assume on* φ *the hypotheses* H_1-H_6 *then each closed set* $\mathfrak{C}^*$ *of rectifiable curves contains a minimizing curve for the functional* λ_φ(C). It seems rather remarkable that hypotheses H_1-H_6 in this theorem need not be formulated with respect to the length but may be assumed with respect to any functional $\lambda(C)$ satisfying condition (1) of §2. This latter is guaranteed for an extended class of functionals $\lambda_\varphi(\varphi)$: e.g., for all curve-integrals along C, satisfying the two following conditions: (α) For each two distinct points q,r the lower bound of the numbers λ(C) *for all curves* C *joining* q *and* r *is* > 0. (β) *Each point* p *of* S *can be joined with each sufficiently neighboring point by a curve* C *for which the diameter and* λ_φ(C) *are arbitrarily small.*

6. *The class* $\mathfrak{C}^*$. $\varphi(p;q,r)$ may be called *seminormal in* p if there exists a neighborhood U and a number $\omega(p) > 0$ such that $\lambda_\varphi(P) > \omega(p)\lambda(P)$ for all[6] closed polygons P of U. We now replace H_5 by the stronger condition

H_5'. φ *is seminormal in each point* p.

Then it can be proved that $\lambda_\varphi(C)$ exists *for each continuous curve* and

that conditions (2′) and (4) of §2 are satisfied. Hence H_6 is equivalent to the hypothesis

H_6'. *For each finite* ρ *the set of all curves* C *with* $\lambda_\varphi(C) \leqq \rho$ *is compact.*

By a method due to Hahn and Tonelli it can be shown, furthermore, that H_6, in this case, is equivalent to

H_6^0. *The lengths of all curves* C *with* $\lambda_\varphi(C) \leqq 0$ *are bounded.*

Therefore we deduce from the last theorem of §2: *If* φ *satisfies conditions* H_1-H_4, H_5' *and either* H_6^0 *or* H_6' *then for each closed set of continuous curves* $\mathfrak{C}^*$ *the set* $\mathfrak{C}_0^*$ *of all rectifiable curves of* $\mathfrak{C}^*$ *contains at least one curve which minimizes* $\lambda_\varphi(C)$ *with respect to all curves of* $\mathfrak{C}^*$ (unless the set $\mathfrak{C}_0^*$ is vacuous).

7. *Non-rectifiable minimizing curves.* Instead of H_5' we now assume the weaker hypothesis: *On each rectifiable curve* C *there may exist a denumerable closed set of points in which* φ *is not seminormal.* If we restrict ourselves to *positive semidefinite* functions $\varphi(p;q,r) \geqq 0$ and assume that φ satisfies condition (β) of §5, then we *still* can prove: Each closed set of continuous curves $\mathfrak{C}^*$ contains one for which $\lambda_\varphi(C)$, or at least an improper length in S_φ, minimizes $\lambda_\varphi(C)$ with respect to all curves of $\mathfrak{C}^*$. Here H_6' is guaranteed if, in particular, $\lambda_\varphi(C)$ satisfies condition (α) of §5. But in these cases *the minimizing curve need not be rectifiable.* In fact, Hahn discovered integrands in the plane for which the minimizing curves are spirals of infinite length, and Carathéodory has shown that in the plane this happens for as simple an integrand as the square-root of the polynomium $(\xi_1^2 + \xi_2^2)(\xi_1'^2 + \xi_2'^2) + 2(\xi_1^2 + \xi_2^2)(\xi_1\xi_2' - \xi_2\xi_1')(\xi_1\xi_1' + \xi_2\xi_2') + (\xi_1\xi_1' + \xi_2\xi_2')$. Our above-mentioned theorem proves the existence of minimizing curves for an extended class of integrands provided that general continuous curves (also non-rectifiable ones) are admitted in the closed set $\mathfrak{C}^*$ of admissible curves.[8]

8. *Vector spaces.* In a vector space the hypotheses H_1-H_5 assume a simpler form if we consider the particular case[9] that $\varphi(p;q,r)$ is *a function of point and direction* $\varphi(p, \delta)$, i.e., that for each point p we have $\varphi(p;q,r) = \varphi(p;q',r')$ provided that the vectors from q to r and from q' to r' are parallel and directed in the same sense. For such a function quasiregularity in p, i.e., the condition $\tau(p) = 0$, means, according to a theorem of Alt,[10] that its indicatrix in the point p is *projectively convex*, i.e., may be transformed into a convex $(n-1)$-dimensional surface by a collineation. This includes Tonelli's quasiregularity in p. In general $\tau(p)$ is a measure for the deviation of the indicatrix in p from projective convexity. For a function φ which in each point p is l.s.c. and quasiregular (i.e., satisfies the conditions $\tau(p) = \sigma(p) = 0$ for each p) the hypothesis H_5' is guaranteed by the following condition: *For each point* p *the lower bound of the numbers* $\varphi(p,\delta) + \varphi(p, -\delta)$ *for all directions* δ *is* > 0, where $-\delta$ denotes the direction opposite to δ. This condition which includes seminormality in Tonelli's sense seems to be as easy to handle as any of the conditions

concerning the integrand in calculus of variations; e.g., in the case of a continuous integrand in the plane it states that $\mathrm{Min}[\varphi(\xi_1, \xi_2; \xi_1', \xi_2') + \varphi(\xi_1, \xi_2; -\xi_1', -\xi_2')] > 0$.

9. So far, our point of view was this: A given primary metric is distorted by a variable multiplicator ("integrand") into a secondary metric. We asked for conditions concerning the integrand such that in the distorted metric all continuous curves (or at least all which are rectifiable in the primary metric) have a length, that this length is l.s.c., and that there exist geodesics. This conception may still be generalized by merely assuming a topological space T as basis without any primary metric at all. The problem, then, is to introduce continuously a metric (in the most general sense) in T in such a way that all (or certain) continuous curves of T have a length, that this length is l.s.c. and that there are geodesics. The metric to be introduced merely consists[11] in the association of a real number with each ordered pair of points and, in particular, of 0 with each pair of identic points. The metric is to be continuous, i.e., if the points $p_1, p_2, \ldots$ converge toward p in T then the numbers associated with the pairs $p_n p$ and $p p_n$ converge toward 0 with $n \longrightarrow \infty$. In the preceding paragraphs a metric of T was associated with each point p of T, viz., the tangential metric $\delta_{\varphi p}(q,r)$ from which the secondary metric can be obtained by the formula $\delta_\varphi(q,r) = \delta_{\varphi q}(q,r)$. The conditions consisted in hypotheses concerning the individual tangential metrics and hypotheses concerning the connection of the tangential metrics in different points with each other. In the classic problems the tangential metrics are obtained from a primary metric by a multiplicative distortion.

[1] Since the generality of a solution depends, so to speak, on four parameters S, φ, $\mathfrak{C}^*$, $\mathfrak{C}_0^*$, it is obvious that not all solutions may be arranged into a linear order of increasing generality.

[2] In this case $\lambda_\varphi(C)$ is, obviously, l.s.c. on $\mathfrak{C}_f$ not only on $\mathfrak{C}_\rho$ for each ρ.

[3] A general report on metric geometry is contained in *L'enseignement mathematique*, **35**, 348–372 (1937). For applications to calculus of variations see C. R. Paris, **201**, p. 706; **202**, p. 1007 and p. 1648, and issue 8 of *Ergebnisse eines mathematischen Kolloquiums*, Vienna, 1937 (in press).

[4] It would be sufficient to assume that there exists a $\rho > 0$ such that for each point p the function $\varphi(p; q, r)$ is defined for all pairs of distinct points q,r at a distance $< \rho$ from p. In the application to vector spaces (§8) this restriction, however, does not make any difference.

[5] In my notes, C. R. Paris,[3] I assumed that for each $\sigma > 0$ the set of all points for which the deviation from *continuity* is $\geqq \sigma$, has the linear measure 0. Dr. Aronszajn in Paris called my attention to the fact that all I actually use in the proofs is the weaker hypothesis H_1 concerning the points with deviation from *lower semicontinuity*. Meanwhile I found in an interesting recent paper by McShane, in the *Duke Journal*, **2**, 597 (1936), the introduction of lower semicontinuity of the integrand. One will notice that we even admit a considerable set of points without lower semicontinuity.

[6] It would be sufficient to assume this for all closed polygons P for which the numbers $\delta(p_i, p_{i+1})$ are sufficiently small.

[7] We replace the curve-integral of φ by the length in S_φ. This length we define in the ordinary way as the limit of the lengths of sub-polygons (cf., my theory of length in metric spaces in *Mathem. Annalen*, 103). In the case of euclidean spaces and functions of point and direction Bouligand proposed to consider the Riemannian curve-integral as limit of its Riemannian sums (*Mém. de la Soc. Roy. d. Sc. de Liége*, 3^{me} sér., t. 19).

[8] One will notice that it is only the application of metric methods which enables us to deal with general continuous curves in calculus of variations. In other examples due to Hahn and Carathéodory the minimizing curves are not even continuous but spirals which approach asymptotically a circle. This seems to suggest another possible extension of the theory in the direction of the admission of certain more general curves.

[9] It seems very likely that in physics or in mathematical economics there will be found applications also of our general case $\varphi(p;q,r)$.

[10] Issue 8 of *Ergebnisse eines mathem. Kolloquiums*, Vienna, 1937 (in press).

[11] Fréchet's concept of metric space was often considered to be too general for useful applications. In fact, it is not general enough for applications to calculus of variations.

ERRATUM

Equation (4) of the paper "On the Mean Value of r^s for Keplerian Systems" in the February, 1937, issue of these PROCEEDINGS should read

$$\overline{r^{-q-2}}/\overline{r^{q-1}} = \left(\frac{2Z}{na_0}\right)^{2q+1} \frac{(2l-q)!}{(2l+q+1)!} \qquad (4)$$

SIMON PASTERNACK